An open dataset of EurekAlert! press releases for science communication research

Zhang, J.; Dudek, J.; Orduña-Malea, E.; Mazoni Fernandes, A.; Costas Comesana, R.

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

An open dataset of EurekAlert! press releases for science communication research

Jingwen Zhang[1]✉, Jonathan Dudek[2], Enrique Orduña-Malea[3], Alysson Fernandes Mazoni[4] & Rodrigo Costas[2,5]

Press releases covering novel scientific developments play an important role in science communication. Their number has increased in the past years, making them a valuable object of quantitative analysis in science communication research. However, large-scale research on academic press releases is limited, likely due to the lack of comprehensive and open datasets. In this *data paper*, we describe how we created a dataset of press releases published on EurekAlert!, an online science news service. The dataset consists of the metadata of a total of 566,566 press releases. We describe how we extracted the data, transformed it into a relational database, and made the code and data openly available to the research community. Next to that, we explain the structure of the dataset and how it can be accessed and propose potential integrations into other research information systems.

Background & Summary

News media play an important role in disseminating scientific publications to general audiences. With the rise of Internet technologies, not only have news sites turned online, but new actors have entered the stage, such as blogs and various Social Media platforms[1]. Still, news media remain one of the most common sources for citizens to learn about scientific developments[2]. Scientific publishers and academic organizations have professionalized the dissemination of science news, e.g., by establishing public information officers (PIOs). The latter send out press releases to inform the public and the press about noteworthy news or events, e.g. a new publication or other scientific news of general interest. As such, the 'academic press release'[3] does not differ much from press releases known from non-academic areas. Significantly, press embargoes help in timing and synchronizing academic press releases across many news outlets: press releases may only be published after an embargo date, but are released to journalists early to give time for preparing their reporting. Platforms like EurekAlert! play an important role as brokers between PIOs and journalists: While the former send press releases to EurekAlert!, the latter get early access to press releases through EurekAlert!.

EurekAlert! (https://www.eurekalert.org) is an editorially independent, non-profit, online science news service, launched and operated by the American Association for the Advancement of Science (AAAS) in 1996. It was established to fill a gap noticed by science journalists, press officers and journal publishers, who wished to use the possibilities of the Internet to send and receive their science research news more broadly[4]. Nowadays, EurekAlert! disseminates news from universities, medical centers, journals, government agencies, and other research organizations[5]. It offers press releases in English, French, German, Spanish, Portuguese, Japanese, and Chinese and has more than 10,000 PIOs and nearly 12,000 journalists registered worldwide in 2016[6]. EurekAlert!'s focus on being an intermediary between journalists and emitters of press releases, providing access to (unredacted) academic press releases at a global scale and across scientific disciplines makes it stand out from other online news services. A comparable service is AlphaGalileo but has fewer (2,000) contributors and journalists (7,000) listed[7]. This makes EurekAlert! a prime source for studying academic press releases at a global scale.

[1]Library of Ningbo University, Ningbo University, Ningbo, China. [2]Centre for Science and Technology Studies (CWTS), Leiden University, Leiden, The Netherlands. [3]The iMetrics Lab. Universitat Politècnica de València, València, Spain. [4]Department of Science and Technology Policy, Institute of Geosciences, University of Campinas, Campinas, Brazil. [5]DSI-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP), Stellenbosch University, Stellenbosch, South Africa. ✉e-mail: zhjingwen1995@163.com

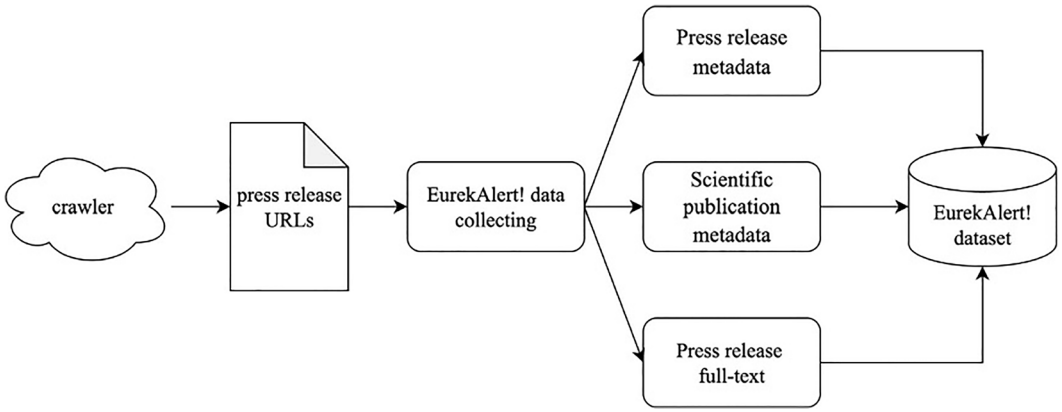

Fig. 1 Data collection workflow of the EurekAlert! dataset.

Linking the general public to science, academic press releases hold interest from the perspective of science communication, both in small and large-scale quantitative analyses, as done, e.g. by Autzen[8] and Sumner *et al.*[9]. Other studies have covered the quality of information in press releases[10] and potential differences with the publications they promote[11]. Academic press releases, and those coming from EurekAlert! in particular, have also become the focus of research in the area of altmetrics, which studies online traces of scientific impact[12]. Bowman and Hassan[13] conducted the first descriptive analysis of EurekAlert! press releases using the Altmetric.com database as a data source, combined with a web-scraping approach. They found that EurekAlert! was the second largest news source on Altmetric.com mentioning scientific publications. Lemke *et al.*[14] identified a potential association between an article's performance and certain qualities (structure, accessibility, and engaging narrative) of its press releases.

Since data on academic press releases is not readily available, various studies have extracted the necessary data every time anew, applying different approaches[13,15]. This hints to a barrier for large-scale data-driven research on academic press releases, including from EurekAlert!: the lack of a systematic summary of data, data structures and research directions. This barrier for reproducibility and the development of quantitative analyses of academic press releases calls for a structured, comprehensive, open, and well-documented database. Such a database would allow researchers to explore new research questions, test hypotheses, and reproduce results. In fact, the lack of such a database may be seen as an important limitation to the introduction and application of large-scale quantitative approaches in the study of science communication processes. While previous work[15] has already presented large-scale analyses of academic press releases, a systematic outline for collecting and structuring the data is still lacking. This paper aims to fill that gap and describes a comprehensive dataset of EurekAlert! press releases[16]. We provide a detailed description of the collection and curation of EurekAlert! press release-metadata and the creation of a relational database for those records, building on the framework discussed by Orduña-Malea and Costas[15].

In presenting a data paper, we add to existing examples in the fields of Scientometrics and Science of Science studies[17–19]; for the part of science communication, we expect our contribution to pave the way for new, particularly quantitative research directions: from descriptive statistics on volume, topics, and contributing organizations, to more advanced analyses linking press releases with scientific publications, social media, and citation data. Moreover, the data supports studying potential biases in science communication, such as overrepresentation of certain topics, as well as changes over time in topical coverage, institutional representation, or geographic origin. Comparisons between the content of scientific publications and their corresponding press releases offer a way of examining accuracy, readability, and framing. When combined with altmetrics and citation data, it also becomes possible to assess the downstream impact of press releases in terms of public attention and academic visibility.

Ultimately, we view this dataset[16] as a starting point for broader efforts in press release-based science communication research. We invite others to build upon this resource, whether by linking it with additional sources (e.g., ROR, PubMed, or Mendeley), expanding it with multilingual press releases, or using it to explore new hypotheses. By publishing an open dataset and the related scripts and procedures, we follow open science principles[20] and aim to facilitate future research for those interested in press releases and science communication dynamics. Ideally, science communication researchers will follow up by producing similar datasets and curation approaches, collectively increasing the analytical realm of science communication research.

Methods

The creation of the dataset[16] started with a web-crawling approach to collect the EurekAlert!-press releases from the URL https://www.eurekalert.org/. The raw data collected was parsed to extract the press release's metadata elements. We will use the term 'metadata' to refer to all the informational elements that describe a press release (e.g. its title, keywords, publication date etc.), but excluding the full-text. To illustrate how data from other (online) sources and research information systems can complement the metadata of the EurekAlert! press releases, we also collected the metadata of scientific publications as it is provided in the records of press releases. Finally, we created a relational database.

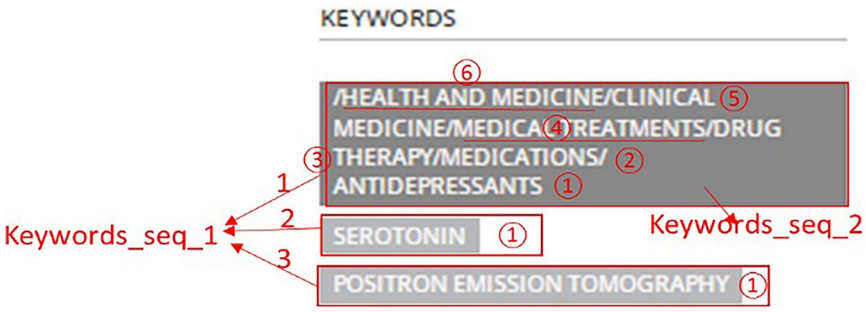

Fig. 2 Example of keyword combinations.

Data collection. The data collection covered three parts: (1) the metadata of the press releases, (2) the metadata of Scientific publication and (3) the full-text of the press releases. Figure 1 shows the data collection and processing flow.

Building on the dataset originally collected by Orduña-Malea and Costas[21], we used SocSciBot v4[22] to retrieve all EurekAlert! press release URLs published between 2021 and 2023 in 2023, and extended the dataset in 2025 to include URLs from 2023 to 2025. All URLs began with 'eurekalert.org/news_release/'. The number at the end of each press release's URL (e.g., https://www.eurekalert.org/news-releases/838589) was used as that press release's unique ID (labelled 'euid'). The content hosted on Eurekalert!'s website was scraped for scientific non-commercial data and text mining (DTM) purposes, following friendly practices (without exceeding one query per second). For each press release, the web crawling returned an HTML file containing the content of the respective web page. Over the course of our research, we crawled the EurekAlert!-website twice: in April 2023 and in April 2025. In April 2025, we collected 566,566 records.

Although we collected EurekAlert! data as comprehensively as possible, a few omissions happened due to missing and broken links. In addition, our data model of EurekAlert! follows what was available at the time of the data collection (March 2025) and does not cover changes of the website and its structure that may have happened in the meantime.

From the data collection by Orduña and Costas[21] in March 2021 to our data collection four years later, 32,667 EurekAlert! press releases became unavailable, which means that some fluctuation in the data has to be accounted for. However, the phenomenon of online records related to scientific publications disappearing over time is not uncommon, at least in the area of altmetrics[23,24].

Data processing. The data collected was first cleaned by removing duplicates. Then, the "|" symbols in the data were replaced with "\" for subsequent saving as a txt file using "|" as a delimiter.

Since we aimed to implement the data in a user-friendly and intuitive way, we created a relational database model[25]. Relational models account for one-to-many relationships between entities. For example, a press release may have more than one keyword or tweet mentioning it. Instead of keeping all the information relating to keywords and tweets together, the relational data model splits it into different tables that can be connected through matching keys, such as unique identifiers and sequence numbers. In the EurekAlert! metadata we have identified the following one-to-many relations: institutions (press releases may report more than one institution), full-text URL links (in the full-text of press releases we usually identify more than one URL linking to external sources or linking to different papers), and keywords (press releases have generally more than one keyword). The relationship between a EurekAlert!-press release and the DOI is one-to-one.

Most EurekAlert! press releases have more than one keyword assigned to them. In the data, we reflect this by randomly assigning them to a sequence: 'keywords_seq_1' (Fig. 2). The keywords on display on the page of a press release are part of a hierarchy that unfolds upon clicking on the keyword. This can be seen in Fig. 2: keywords with the 'keywords_sequence_1'-numbers 2 and 3 are not unfolded yet, while number 1 is. 'Keywords_seq_2' reflects the hierarchy behind each keyword: '1' is the sequence number for the keyword at the lowest level, '6' in this example stands for the highest level. The hierarchy of keywords can be up to 11 levels deep.

Data depositing. We made the dataset openly available on DataverseNL[16], an institutional research data repository. It is deposited with a CC BY-NC 4.0 license, following a consultation with and the written permission of EurekAlert!. This license excludes commercial (re-)use, requires proper attribution of the source, and an indication of whether any changes have been made (https://creativecommons.org/licenses/by-nc/4.0/).

Data Records

Data files and supporting materials. The dataset is available at DataverseNL[16], and can be accessed as a JSON file within the **EurekAlert_dataset_2025.zip** archive. This file contains all metadata collected from the EurekAlert! platform and can be parsed into different relational tables, following the structure presented in Fig. 4. These relational tables can be used in applications such as Microsoft Access, MS SQL Server, Google BigQuery, and similar relational database systems.

In addition to the dataset, the Dataverse record also includes supporting code and documentation files to facilitate reuse and integration: **(1) EurekAlert-in-json_codeScripts.zip**,including **Eurekalert to BigQuery**

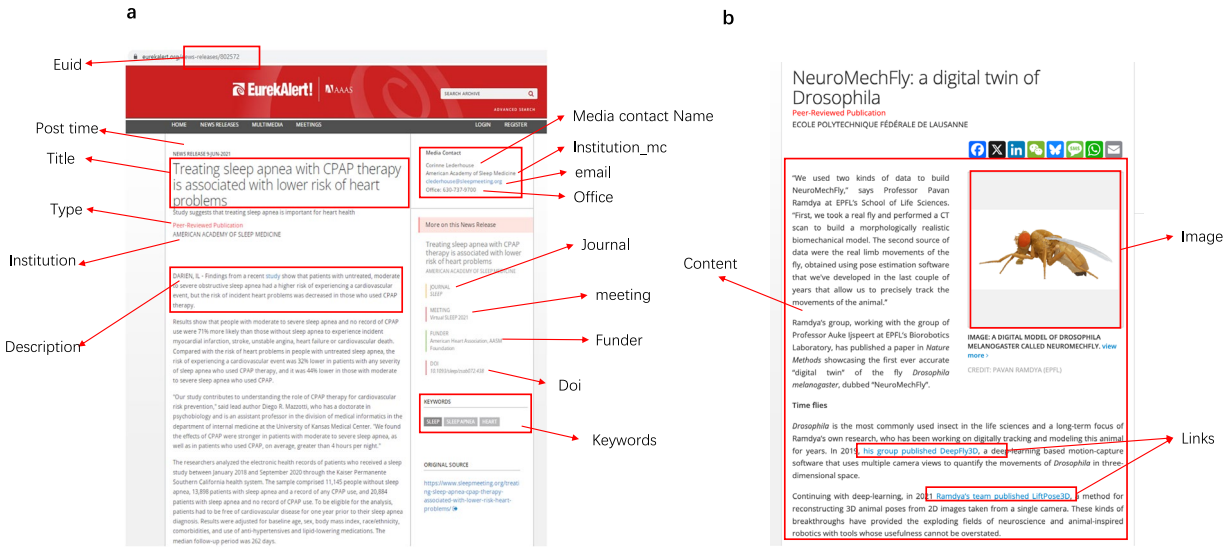

Fig. 3 Screenshot of press release page. (**a**) Eurekalert! press release structure (front-end version); (**b**) Full-text components of EurekAlert! press releases.

Data	Field name	Description
1) Press release metadata	Euid	Unique identifier of each press release
	URL	press release URL
	Title	Main title of the press release
	Post-time	The post-time of the press release, indicated in day-month-year format
	Type	Type of press release, including Book Announcement, Business Announcement, Grant and Award Announcement, Meeting Announcement, Peer-Reviewed Publication and Reports and Proceedings.
	Institution	The institution emitting the press release. Usually a university, a scientific journal, etc.
	Description	A short text of about 300–500 characters describing the main topic of the press release
	Keywords	Keywords describing the scientific topics of the press release
	Original source	Website on which the press release has been published originally
2) Scientific publication metadata	Journal	The journal of the scientific publication covered
	Meeting	The scientific meeting or conference that published the scientific publication covered
	Funder	Information about the funder(s) supporting the scientific publication
	DOI	Digital Object Identifier of the scientific publication covered
3) Full-text data	Links	The hyperlinks (links) mentioned in the full-text

Table 1. Data fields extracted for the EurekAlert! press release records. Note: (Optional) means that some metadata categories or fields are not available for all press releases.

2025.ipynb, a notebook illustrating how to read the JSON file and create corresponding tables in Google BigQuery; **data_clean.sql**, a script for initial data cleaning and transformation; **eurekalert_data_processing.sql**, a script for creating structured tables from raw JSON fields; **eurekalert_metadata&fulltext.ipynb**, the core notebook for crawling, parsing, and exploring both metadata and full-text content from the EurekAlert! platform; **(2) README.txt**, containing detailed descriptions of the dataset fields, structure, and usage instructions.

Description of dataset columns. The various metadata fields to be extracted for each press release were identified by examining the web-scraped HTML-files of the press releases. The metadata fields extracted from each press release, including the bibliographic data of the scholarly publication(s) mentioned, are depicted in Fig. 3(a). On the right side of the press release record, the information of the scientific publications being promoted are provided, such as the journal, meeting, funder and DOI. To enable the full-text analysis of press releases, the full-text of each EurekAlert! press release was parsed. This includes the text, images, and links, see Fig. 3(b).

Combining the data extracted from each press release record, we can distinguish four categories: 1) press release metadata, 2) scientific publication metadata, 3) data extracted from the press releases' text parts. Table 1 lists all the data fields and their descriptions per category.

Construction of the dataset. The relational model of the final dataset[16] is shown in Fig. 4. All related tables are connected by the euid as the main identifier and contain information as specified in Table 1. A press release

Version	Collection time	Total Records	Inaccessible Records	Accessible Records	Coverage period	New records Added
2025 version	Apr 2025	566,566	1 (vs. 2023)	495,179	Jan 1996-Mar 2025	71,368
2023 version	Apr 2023	495,180	32,552 (vs. 2021)	422,227	Jan 1996-Mar 2023	72,953
2021 version	Mar 2021	454,779	—	—	Jan 1996-Feb 2021	—

Table 2. Comparison of the three dataset versions in terms of data retrievability and coverage.

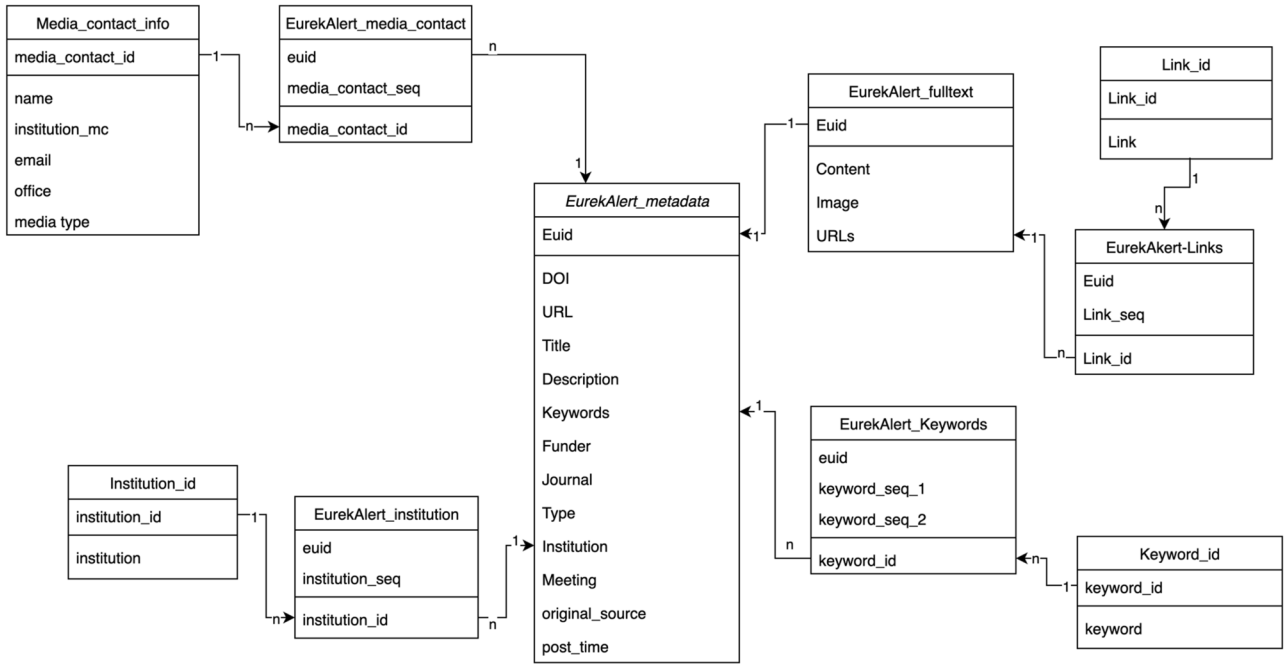

Fig. 4 The relational model of the EurekAlert! dataset.

may involve multiple keywords, institutions, and Links. To avoid data duplication, reduce redundancy, and make the data consistent, all one-to-many metadata fields were split and assigned IDs to data in secondary tables (e.g. 'EurekAlert_Keywords', 'EurekAlert_Institution', and 'Keyword_id', 'Institution_id', respectively), then connecting them to the press releases via the euid, as shown in Fig. 4.

This relational model design provides a clear structure that is easily comprehensible. The various ids and sequence numbers can be specified as keys, improving query performance. Modifications of and additions to the data can happen without having to restructure the entire dataset.

Technical Validation

Data completeness over time. To evaluate the reliability and completeness of our data collection approach, we compared three versions collected at different points in time: 2021, 2023, and 2025. The version from 2021 was collected by Orduña-Malea and Costas[21] and served as initial reference. Table 2 shows that the total number of new records increased with every update. This growth was relatively consistent with about 70,000 additional press releases each time. Each version contained records dating back to January 1996, the starting year of EurekAlert!. All this demonstrates both the technical robustness of our data collection approach and the consistency of the data source.

Between the initial (2021) and the 2023 version, a significant change occurred on the EurekAlert!-platform. In the earlier version, each news release was identified solely by its URL, which we used as a unique identifier. In the newer version, EurekAlert! changed the URLs of press releases to include a unique identifier (see Data Collection in Methods). In addition, the webpage layout and underlying HTML structure were updated, which resulted in previously accessible URLs becoming inaccessible. These changes directly contributed to a considerable number of records becoming inaccessible in the 2023 version (n = 32,552, 7.1%). In response, the data collection process was fully reengineered to accommodate the new ID-based structure, ensuring compatibility with the updated platform. The effectiveness of this adaptation is reflected in the 2025 dataset's[16] near-perfect retrieval success rate.

These results not only confirm the reliability of EurekAlert! as a long-term data source but also demonstrate the technical robustness and adaptability of our collection methodology in the face of dynamic web environments.

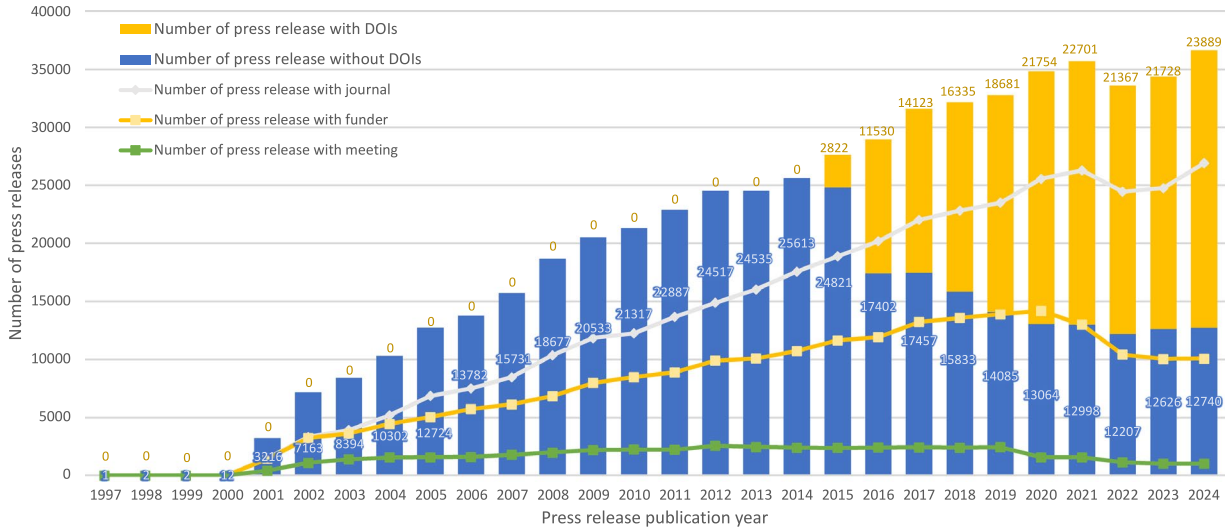

Fig. 5 EurekAlert! press releases per year from 1997 to 2024.

Completeness of scientific publication metadata. To connect press releases to scientific publications, complete information on the journals, funding, and the DOIs of the scientific publications reported on is essential. Figure 5 shows that key metadata fields (journal, funder, and meeting) have steadily grown over time, with journal information in particular. This growth matched the overall increase of the annual volume of press releases.

Prior to 2015, EurekAlert! press releases did not indicate DOIs in the dedicated field in the right margin of press releases' web pages. Since 2015, the number of press releases with DOIs has grown consistently, with more than 50% of the press releases having a DOI from 2018 onwards. It must be noted that EurekAlert! also refers to pre-publication articles[12], potentially resulting in some research articles reported on to not having a DOI at the time of the news release yet. These findings show that a considerable part of the press releases in EurekAlert! can be linked to bibliometric metadata, with increasing shares in more recent years.

Limitations. Despite our efforts to collect the EurekAlert! data as comprehensively as possible, the dataset[16] has several limitations that users should be aware of when reusing the data. First, there are minor omissions due to broken or missing links at the time of data extraction. Additionally, our data model was developed based on the structure of the EurekAlert! platform as it existed during data collection and may not fully reflect changes introduced on the website afterwards.

The only means for connecting press releases to scholarly outputs in an unambiguous, ready way is through the DOIs that are provided by a subset of the press releases. This means that not all the publications that could be linked to a press release are covered. While text-mining techniques could be applied to infer mentions of publications, direct identifier matching remains more reliable and was prioritized. For press releases without DOIs, we have not yet attempted large-scale matching to OpenAlex or other databases. Such linkage would require approximate or heuristic methods using information such as titles, journal names, authors, or dates. This approach is technically feasible but outside the scope of the current Data Descriptor paper.

Another important limitation is that name disambiguation for entities such as institutions, authors, and journals was not performed. While we applied basic data cleaning procedures such as removing duplicates, trimming extra whitespace, and standardizing simple variations in capitalization or punctuation, no comprehensive normalization or disambiguation was carried out. The names included in the dataset reflect the original values as displayed on the EurekAlert! platform and may contain inconsistencies due to variations in spelling, punctuation, or formatting. Further disambiguation, particularly of institutional names, would significantly enhance the analytical precision and interoperability of the dataset. However, this is a complex, domain-specific task that typically requires the use of external authority files (e.g., ROR) or advanced algorithms and is beyond the scope of the current Data Descriptor. We identify this as a promising area for future work.

Another limitation of our dataset is that it does not contain the full-text of the press releases due to copyright reasons. Researchers interested in obtaining (subsets of) the full-text may adapt our code to extract them from the EurekAlert! website.

Finally, the dataset primarily consists of English-language press releases. While EurekAlert! offers content in other languages (e.g., Chinese, Japanese, Spanish), these were not included in this release. Future versions of the dataset may expand to include multilingual content to support cross-cultural studies in science communication.

Usage Notes

Accessing and using the dataset on google BigQuery. Next to accessing the dataset on DataverseNL[16], it is also publicly available on Google BigQuery (GBQ) as part of the InSySPo project at the State University of Campinas, Brazil[26]. With this version, users can simply run queries directly in the cloud without the need to download the dataset and create a database. All press release metadata described in this paper is included. Users

ID	keyword	number of records
1	Health and medicine	238801
2	Life sciences	238465
3	Scientific community	174860
4	Social sciences	165426
5	Applied sciences and engineering	150279
6	Physical sciences	135414
7	Diseases and disorders	133443
8	Organismal biology	102206
9	Research programs	89004
10	Clinical medicine	83139
11	Anatomy	66364
12	Medical specialties	65871
13	Physics	54373
14	Psychological science	53776
15	Medical treatments	50536
16	Engineering	50312
17	Education	46812
18	Biochemistry	46482
19	Chemistry	46071
20	Health and medicine	238801

Table 3. Top 20 most-used keywords in EurekAlert! press releases.

can query the data using the GBQ SQL syntax and combine it with other datasets hosted on GBQ or external data sources through joins or federated queries.

To access the dataset,one must have a Google Account with a valid Google Cloud project that is enabled for billing (for query-related charges). The project's address is: https://console.cloud.google.com/bigquery?project=insyspo. The 'insyspo'-project in GBQ comprises multiple public datasets, from which 'publicdb_eurekalert_2025' needs to be chosen. It can also be accessed directly through the following link: https://console.cloud.google.com/bigquery?project=insyspo&ws=!1m4!1m3!3m2!1sinsyspo!2spublicdb_eurekalert_2025. We recommend users new to GBQ to consult Google's official documentation to learn how to set up projects, manage billing, and write queries.

Keyword structure and clustering possibilities. The EurekAlert! dataset[16] includes a rich set of keyword metadata, which offers opportunities for thematic analysis and the classification of press releases. Most EurekAlert! press releases have at least one keyword group, each containing multiple keywords, with a hierarchical relationship. Some of the EurekAlert! press releases contain up to 100 keywords. The top 20 most used keywords in EurekAlert! press releases across all levels in the hierarchy are shown in Table 3. Health and medicine is the most frequently mentioned keyword on EurekAlert!, with almost half of all press releases related to *Life Sciences* and *Health and medicine*. *Scientific Community* and *Social Sciences* are the two following most common keywords. This shows a strong orientation towards disseminating research from the Life and health sciences as well as Social Sciences. This aligns with previous studies in the field of altmetrics which also show that the disciplines of Medical, Health and Social sciences are more often picked up in media and social media platforms[27,28].

This keyword structure enables multiple use cases,such as thematic clustering of press releases using keyword hierarchies; trend analysis of topics over time (e.g., changes in frequency of health-related keywords); discipline-specific filtering for focused analyses (e.g., only social science-related press releases); mapping topical biases in science communication, such as the overrepresentation of specific fields.

To illustrate the structure and co-occurrence of keywords, a co-occurrence map is presented in Fig. 6. In this map, nodes represent keywords, while edges connect keywords that appear together in EurekAlert! press releases (see Van Eck[29] for an explanation of the clustering technique). In the map, we can identify clusters of keywords and gain insights into the underlying structure of keywords in EurekAlert! press releases. For example, there are 7 clusters identified in the co-occurrence map clustering keywords, health and medicine and life sciences (green cluster), organismal biology (purple cluster), scientific community and social sciences (yellow cluster), physical sciences and engineering (red cluster) and environmental sciences (blue cluster).

Notably, the network reveals the hierarchical order of the keywords as well: larger nodes (such as "Life Sciences") correspond to high-level, comprehensive categories that serve as central topics for press releases. Conversely, smaller nodes represent more specialized terms that typically co-occur within these broader thematic frameworks. These clusters highlight the latent structure of press release topics and offer a foundation for further topic modeling and cross-domain comparison in science communication research.

Open dataset of EurekAlert! press releases. The dataset[16] described in this paper can be expanded further by interlinking the metadata elements with other open research information systems, such as Crossref (https://www.crossref.org/), PubMed (https://pubmed.ncbi.nlm.nih.gov/) or OpenAlex (https://openalex.

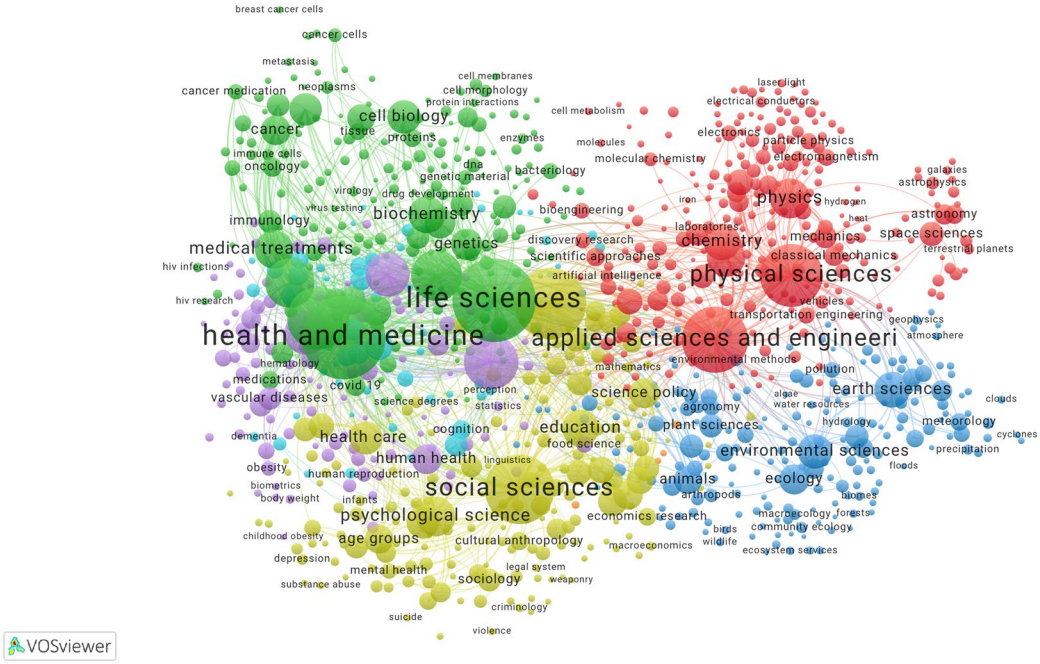

Fig. 6 Co-occurrence network of keywords in EurekAlert!. Explore online: https://app.vosviewer.com/?json=https://drive.google.com/uc?id=1oQJ_4-glqbL4UF5uu4gASu5DibbY4oab.

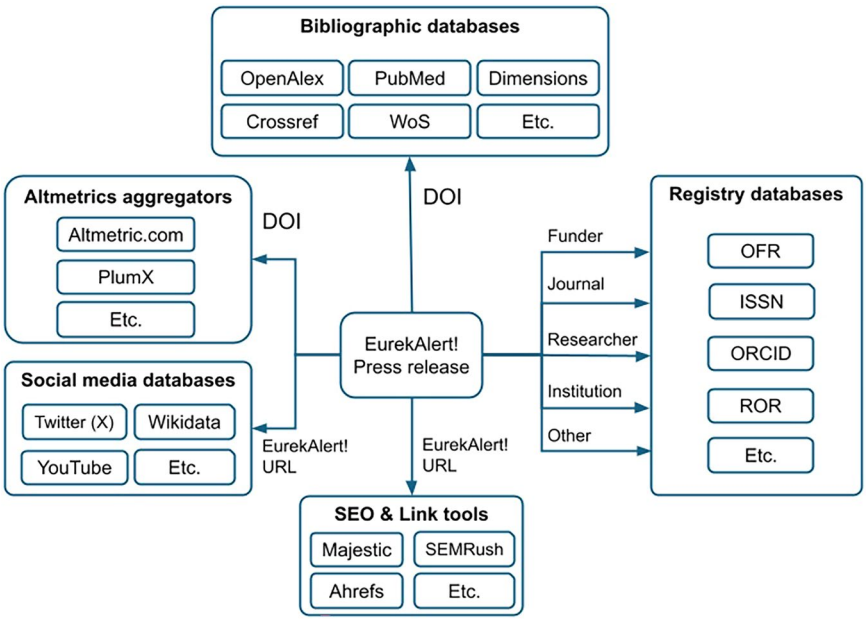

Fig. 7 Open dataset of EurekAlert! press releases.

org/). Here, interconnections could be created by using DOIs or other scholarly publication identifiers. Journals reported in press releases could be linked via ISSNs (https://portal.issn.org/), researchers could be linked via ORCID iDs (https://orcid.org/), funders to, e.g., the Open Funder Registry (OFR) (https://www.crossref.org/services/funder-registry/), affiliations and research organizations to the Research Organization Registry (ROR - https://ror.org/), EurekAlert! keywords could be linked to Wikidata (in the same fashion as "Concepts" have been assigned to publications in OpenAlex - https://docs.openalex.org/api-entities/concepts), and any URL information extracted from EurekAlert! press releases could be connected to backlink information services, similar to the work of Orduña-Malea[30]. Moreover, press releases (their URLs and mentioned DOIs) could be connected to altmetric services like Crossref Event Data (https://www.crossref.org/services/event-data/) to collect further information on other online dissemination activities surrounding press releases. Commercial data sources like Altmetric.com, Web of Science, Scopus or Dimensions could be connected with the EurekAlert! dataset[16] as well.

All these data sources taken together could be described as a knowledge graph with EurekAlert!-metadata at the centre (Fig. 7). Such a knowledge graph may eventually become part of even larger infrastructures, such as the Open Research Knowledge Graph[31].

Code availability

All scripts for collecting the data and creating the relational database in Python MS SQL Server can be found alongside our dataset deposited on DataverseNL[16].

Received: 1 April 2025; Accepted: 20 August 2025;
Published online: 30 September 2025

References

1. Dudo, A. Scientists, the Media, and the Public Communication of Science. *Sociology Compass* **9**, 761–775, https://doi.org/10.1111/SOC4.12298 (2015).
2. Funk, C., Gottfried, J. & Mitchell, A. Science News and Information Today. *Pew Research Center* https://www.journalism.org/2017/09/20/science-news-and-information-today/ (2017).
3. Sumner, P. *et al*. The Association between Exaggeration in Health Related Science News and Academic Press Releases: Retrospective Observational Study. *BMJ* **349**, g7015, https://doi.org/10.1136/bmj.g7015 (2014).
4. Stockton, N. Inside EurekAlert, the News Hub That Shapes the Science You Read. *Wired*. https://www.wired.com/2016/05/internet-hub-science-news-shaping-world-20-years/ (2016).
5. Eurekalert! Using EurekAlert! As a Public Information Officer (PIO). *Eurekalert!* https://www.eurekalert.org/help (n.d).
6. Lane, E. EurekAlert! Celebrates 20 Years at the Forefront of Science Communication. *AAAS* https://www.aaas.org/news/eurekalert-celebrates-20-years-forefront-science-communication (2016).
7. AlphaGalileo. About us. *AlphaGalileo* https://www.alphagalileo.org/en-gb/AlphaGalileo/About-us (n.d).
8. Autzen, C. Press releases—the new trend in science communication. *J. Sci. Commun.* **13**(3), C02, https://doi.org/10.22323/2.13030302 (2014).
9. Sumner, P. *et al*. Exaggerations and Caveats in Press Releases and Health-Related Science News. *PLoS ONE* **11**(12), e0168217, https://doi.org/10.1371/journal.pone.0168217 (2016).
10. Choi, A. R. & Feller, E. R. Misrepresentation of mild traumatic brain injury research in press releases. *PM&R* **14**(7), 769–778, https://doi.org/10.1002/pmrj.12656 (2022).
11. Mellor, D. D. & Green, D. J. A critical review exploring science communication of nutrition and dietetic research: a case-based approach exploring methodologies. *J. Hum. Nutr. Diet.* **36**, 1468–1479, https://doi.org/10.1111/jhn.13155 (2023).
12. Priem, J., Taraborelli, D., Groth, P. & Neylon, C. Altmetrics: A manifesto. *altmetrics.org* http://altmetrics.org/manifesto (2010).
13. Bowman, T. D. & Hassan, S. Science news and altmetrics: looking at EurekAlert! *In The 2019 Altmetrics Workshop*. BYOR: Bring Your Own Research (Stirling, UK) http://altmetrics.org/wp-content/uploads/2019/10/Bowman_altmetrics19_paper_6.pdf (2019).
14. Lemke, S., Sakmann, J., Brede, M. & Peters, I. Exploring the relationship between qualities of press releases to research articles and the articles' impact. *In Proc. 2021 Int. Conf. Scientometrics Informetrics*. 639–644, https://www.researchgate.net/profile/Steffen-Lemke-2/publication/356471196_Exploring_the_Relationship_between_Qualities_of_Press_Releases_to_Research_Articles_and_the_Articles'_Impact/links/619d14f507be5f31b7aeb8e2/Exploring-the-Relationship-between-Qualities-of-Press-Releases-to-Research-Articles-and-the-Articles-Impact.pdf (2021).
15. Orduña-Malea, E. & Costas, R. 1. A scientometric-inspired framework to analyze EurekAlert! press releases. In *The Science-Media Interface* (eds. Broer, I. *et al*.) 1–26, https://doi.org/10.1515/9783110776546-001 (De Gruyter Saur, 2023).
16. Zhang, J., Dudek J., Mazoni A., Orduna-Malea E. & Costas R. EurekAlert!: An open dataset for science communication research, *DataverseNL* https://doi.org/10.34894/EZO4JE (2025).
17. Waltman, L. & Larivière, V. Special issue on bibliographic data sources. *Quantitative Science Studies* **1**(1), 360–362, https://doi.org/10.1162/qss_e_00026 (2020).
18. Lin, Z., Yin, Y., Liu, L. & Wang, D. SciSciNet: A large-scale open data lake for the science of science research. *Scientific Data* **10**(1), 315, https://doi.org/10.1038/s41597-023-02198-9 (2023).
19. Arroyo-Machado, W., Torres-Salinas, D. & Costas, R. Wikinformetrics: Construction and description of an open Wikipedia knowledge graph data set for informetric purposes. *Quantitative Science Studies*. 1–22, https://doi.org/10.1162/qss_a_00226 (2022).
20. UNESCO. UNESCO Recommendation on Open Science, https://doi.org/10.54677/MNMH8546 (2021).
21. Orduña Malea, E. & Costas, R. The Eurekalert! project: dataset of mentions to press releases. *Univ. Politècnica de València* https://doi.org/10.4995/Dataset/10251/186769 (2022).
22. Thelwall, M. Web Crawling: SocSciBot. *In Introduction to Webometrics: Quantitative Web Research for the Social Sciences*. Synthesis Lectures on Information Concepts, Retrieval, and Services, https://doi.org/10.1007/978-3-031-02261-6_6 (Springer, Cham, 2009).
23. Arroyo-Machado, W. & Torres-Salinas, D. Stranger things: the vanishing of the Altmetric Attention Score values in information and library science. *Scientometrics* **129**, 6287–6300, https://doi.org/10.1007/s11192-024-05011-5 (2024).
24. Fang, Z., Dudek, J. & Costas, R. Facing the volatility of tweets in altmetric research. *J. Assoc. Inf. Sci. Technol.* **73**(8), 1192–1195, https://doi.org/10.1002/asi.24624 (2022).
25. Date, C. J. An Introduction to Database Systems, 8th edn, 59–81 (Addison Wesley, 2004).
26. Mazoni, A. & Costas, R. Towards the democratisation of open research information for scientometrics and science policy: the Campinas experience. *Leiden Madtrics* https://doi.org/10.59350/eqmfk-82y98 (2024).
27. Costas, R., Zahedi, Z. & Wouters, P. The thematic orientation of publications mentioned on social media: Large-scale disciplinary comparison of social media metrics with citations. *Aslib J. Inf. Manag.* **67**(3), 260–288, https://doi.org/10.1108/AJIM-12-2014-0173 (2015).
28. Fang, Z., Costas, R., Tian, W., Wang, X. & Wouters, P. An extensive analysis of the presence of altmetric data for Web of Science publications across subject fields and research topics. *Scientometrics* **124**(3), 2519–2549, https://doi.org/10.1007/s11192-020-03564-9 (2020).
29. Van Eck, N. & Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* **84**(2), 523–538, https://doi.org/10.1007/s11192-009-0146-3 (2010).
30. Orduña-Malea, E. & Costas, R. Link-based approach to study scientific software usage: The case of VOSviewer. *Scientometrics* **126**(9), 8153–8186, https://doi.org/10.1007/s11192-021-04082-y (2021).
31. Jaradeh, M. Y. *et al*. Open Research Knowledge Graph: Next generation infrastructure for semantic scholarly knowledge. *In Proc. 10th Int. Conf. Knowledge Capture*. 243–246, https://doi.org/10.1145/3360901.3364435 (2019).

Acknowledgements
We would like to thank Brian Lin, Director, Editorial Content Strategy at EurekAlert! for providing advice on our work with press releases from EurekAlert! and approving our data availability approach. We also thank Andrew Hoffman, data steward at Leiden University, for advising us on the process of openly publishing the dataset. Rodrigo Costas and Jonathan Dudek are partially funded by the South African DSI-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP). Jingwen Zhang is financially supported by the China Scholarship Council (Grant No.202206190142). Enrique Orduña-Malea is partially funded by PID2022-142569NA-I00, granted by MCIN/AEI/ 10.13039/501100011033 and by "ERDF A way of making Europe."

Author contributions
Conceptualization: all. Data curation: J.Z., E.O., A.M. Validation: J.Z., E.O. Software: J.Z. Investigation: J.Z., J.D., R.C. Writing – original draft: J. Z., J.D. Writing – review & editing: all.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to J.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.