Galileo, a data platform for viewing news on social networks

Luis Cárcamo-Ulloa; Claudia Mellado; Carlos Blaña-Romero; Diego Sáez-Trumper

Nota: Este artículo se puede leer en español en:
https://revista.profesionaldelainformacion.com/index.php/EPI/article/view/86941

Recommended citation::

Cárcamo-Ulloa, Luis; **Mellado, Claudia**; **Blaña-Romero, Carlos**; **Sáez-Trumper, Diego** (2022). "*Galileo*, a data platform for viewing news on social networks". *Profesional de la información*, v. 31, n. 5, e310512.

https://doi.org/10.3145/epi.2022.sep.12

Manuscript received on March 14th 2022
Accepted on July 4th 2022

Luis Cárcamo-Ulloa ✉
https://orcid.org/0000-0003-0633-9606

Universidad Austral de Chile
Instituto de Comunicación Social
Campus Isla Teja
Independencia 641, Valdivia, Chile
lcarcamo@uach.cl

Claudia Mellado
https://orcid.org/0000-0002-9281-1526

Pontificia Universidad Católica de Valparaíso
Escuela de Periodismo
Avenida Universidad 330
Campus Curauma, Valparaíso, Chile
claudia.mellado@pucv.cl

Carlos Blaña-Romero
https://orcid.org/0000-0001-6987-5567

Universidad Austral de Chile
Instituto de Informática
Campus Miraflores
Independencia 641, Valdivia, Chile
carlos.blana@uach.cl

Diego Sáez-Trumper
https://orcid.org/0000-0002-7679-5423

Fundación Wikimedia
Universitat Pompeu Fabra
Barcelona, Spain
dsaez-trumper@acm.org

Abstract

This article aims to introduce *Galileo*, a platform for extracting and organizing news media data on social networks. *Galileo* integrates publications made on the main social networks used in the information ecosystem, namely *Facebook*, *Twitter*, and *Instagram*. Currently, the system includes 97 media outlets from nine countries: Brazil, Chile, Germany, Japan, Mexico, South Korea, Spain, United Kingdom, and United States. *Galileo* uses a *Twitter* API and the service *Crowd-Tangle* to download *Facebook* and *Instagram* posts. This data is stored in a local database and can be accessed through a user-friendly interface, which allows for the analysis of different characteristics of the posts, such as their text, source popularity, and temporal dimension. *Galileo* is a tool for researchers interested in understanding news cycles and analyzing news content on social networks.

Keywords

News; Visualisation; Data science; Textual data; Journalism; Social media; Social networks; Platforms; *Twitter*; *Facebook*; *Instagram*; *Galileo.*

Funding

Researchers Luis Cárcamo-Ulloa and Claudia Mellado have been beneficiaries of the *National Agency for Research and Development* [*Agencia Nacional de Investigación y Desarrollo*] (*ANID-Chile*). In particular, the *ANID-Covid 0172* and *ANID-Plu 210013* initiatives.

1. Introduction

Watts (2016; 2017) highlights the importance of computational social sciences (CSC) in investigating complex phenomena based on vast volumes of data. Every day, the media generate huge amounts of data that reflect the political, economic, and cultural activities of the societies into which they are integrated. For this reason, analyzing the large volume of textual data produced daily by the media is important to understand the biases and biased framing that can harm democracies (**Watts**; **Rothschild**; **Mobius**, 2021).

Within journalism studies, most of the textual analyses conducted for quantitative academic research are obtained through manual quantitative content analyses. Despite the technological revolution and the new digital ecosystem, automatic text coding, especially for complex measurements and latent content, is a relatively new field (**Hamborg**; **Donnay**; **Gipp**, 2019). Enthusiasts and critics alike have developed different arguments to underscore the advantages and disadvantages of automatic text analysis. Some have argued that automatic methods are very promising, fast, and objective (**Pereira** *et al.*, 2015); in contrast, others emphasize that, for example, the analysis of news framing requires a semantic context and intertextual clues beyond the news (**Baden**, 2018). These latter aspects still seem to be very unwieldy for artificial intelligence.

However, overall, various studies agree that the enormous quantity of existing information requires tools that make the expeditious collection, processing, and analysis of information possible (**Grimmer**; **Stewart**, 2013; **Lewis**; **Zamith**; **Hermida**, 2013; **Matthes**; **Kohring**, 2008; **Trilling**; **Jonkman**, 2018; **Van-Atteveldt**; **Peng**, 2018).

Developments in monitoring interactions in information ecosystems are varied and are very efficiently used, for example, for digital marketing purposes. However, it is still difficult for communication professionals and analysts to compile what the media are saying about a topic on social networks and/or on the web in a comprehensive, organized, and efficient way (**Cárcamo-Ulloa** *et al.*, 2017). Normally, obtaining data requires specific developments or scripts to extract the data, depending on the objectives of the analysis (**Zhang**; **Boons**; **Batista-Navarro**, 2019)

Galileo is a platform developed within the framework of the *ANID-COVID 0172* project "Analysis and automatic monitoring of the role of journalism and the media on their different platforms during the phases of the health crisis caused by COVID19 in nine countries in America, Europe and Asia" ["*Análisis y monitoreo automático del rol del periodismo y los medios en sus diferentes plataformas durante las fases de la crisis sanitaria provocada por el COVID19 en nueve países de América, Europa y Asia*"]. This initiative was funded by the *National Agency for Research and Development* of the *Government of Chile*. It should be noted that the *ANID-COVID 0172* project includes communication and journalism researchers from nine countries: Brazil, Chile, Germany, Japan, Mexico, South Korea, Spain, United Kingdom and United States.

The main objective of the *ANID-COVID* project is to design and implement strategies for analyzing and monitoring the media attention cycles on traditional platforms and social networks during the coverage of COVID-19 and evaluate the role that journalism plays in the communication of health crises. In this context, *Galileo* is not limited only to information about the health crisis; rather it includes all the news production that a media group published on social networks from the beginning of the pandemic (January 2020) onwards.
http://www.galileo-jrp.org

This platform grew out of the need to create a tool that allows communication researchers to access the content that media outlets post on their social media accounts, especially considering that platforms such as *Twitter* or *Facebook* account for a large part of news readership (**Salazar**, 2019; **Newman** *et al*., 2021).

As a source of data extraction, media outlets' social networks vary from their websites. In this regard, it is important to note that:

- The amount of textual data collected may change depending on the organizational culture of the press teams or the definition of publishing business[1] considering that the media may not dump all their journalistic content on social networks; however, we can assume with some certainty that they will post the most relevant material.
- News texts on social networks are obviously more concise than the extended news articles the media publish in web or traditional forms. Informative posts on social networks are more similar to journalistic "leads." However, they must also capture at least the basic questions of "what happened," "where," and "when" the events narrated occurred. These traditional structures are indispensable when composing a news item and are answered at the beginning of any informative text.

1.1. Interest in textual data

On social networks, an enormous amount of information circulates, and it does so in compact formats that are consumed by users. The post is the maximization of the summary of information that effective informative journalism has always pursued, and at the same time, it is a new discursive genre that the media display their social networks (**Cárdenas-Neira**, 2016; **Raimondo-Anselmino**; **Sambrana**; **Cardoso**, 2017). **Ojo** and **Heravi** (2018) aimed to analyze patterns and typologies of successful news narratives that could guide the future of data journalism. **Newman**, **Dutton**, and **Blank** (2019) went so far as to propose that the expansion of the news scene into social networks would allow –as online information increases and individuals habitually connect to the Internet– the emergence of a "fifth power" based on the activities

> Analyzing the large volume of textual data produced daily by the media is important to understand the biases and biased framing that can harm democracies

of networked individuals who obtain and disseminate information. On the other hand, in concordance with the filter bubble concept proposed by **Pariser** (2011), it is understood that *Facebook*'s algorithm, for example, is less likely to provide individuals with media posts that run contrary to the users' attitudes (**Levy**, 2021); therefore, the social network algorithms, by limiting exposure to said contrary news items, increase polarization. All this has led us to think that the news that circulates on social networks would be a research topic of particular interest.

Galileo focuses on textual data rather than on user interactions. It is a tool that allows you to manage considerable volumes of textual corpora to carry out manual content analysis, as well as apply automatic language processing techniques for topic modeling, word embedding, and entity analysis (sources and actors).

The *Galileo* platform seeks to facilitate data exploration by testing textual data analysis models to better understand of information ecosystems on social networks. Data science can provide tools for understanding journalism and the media (**Hamborg**; **Donnay**; **Gipp**, 2019). The central challenge for data science applied to the analysis of press reports is then to develop techniques that improve the evaluation of the quality of news in a world that is characterized by growing informational entropy (**Cardon**, 2018).

There are relevant studies in the computerized analysis of great volumes of data on the web and in the media. Thus, computerized techniques have made it possible, for example, to recognize a series of biases on the web, in general (**Baeza-Yates**, 2018), and around political figures that appear in the press, in particular (**Sáez-Trumper**; **Castillo**; **Lalmas**, 2013). Recently, progress has even been made toward the identification of subtle patterns that determine biases in journalistic framing (**Morstatter** *et al*., 2018). As explained by **Schmitz-Weiss** *et al.* (2017), technology has eliminated the barriers to journalistic activity, making it possible to study and analyze information generated by the news itself.

Advances in computational linguistics and automatic language processing can be applied to textual data from news items. Today we are trying to improve the detection of fake news (**Zhou** *et al*., 2019) and the quality of information (**Romanou** *et al.*, 2020). With this in mind, with the textual data extracted and organized by the *Galileo* platform, both traditional content analysis and automated processing can be conducted.

In the literature, we found studies that identified 37 Chilean media outlets within the media corpora (**Vernier**; **Cárcamo-Ulloa**; **Scheihing-García**, 2017). **Jiang** *et al*. (2017), for their part, compare the attitudes of the British media toward climate change over a period of 10 years. The research corpus of this latest work was composed of 11,720 newspaper articles collected between 2007 and 2016 from four UK newspapers (*The Guardian*, *The Times*, *The Telegraph*, and *The Independent*). This work jointly addresses: sentiment analysis and latent Dirichlet allocation (LDA) to identify topics (**Blei**; **Ng**; **Jordan**, 2003). In particular, topic analysis is a fairly widespread strategy when handling large volumes of textual data from the press (**Li** *et al.*, 2020).

1.2. Some tools to access world news

There are good tools for accessing news corpora and studying media content. As examples, we will highlight three platforms.

- *Media Cloud* is a platform for analyzing mass media and has three fundamental tools: *Explorer*, *Topic Mapper*, and *Source Manager*. *Media Cloud* includes a database of sources from more than 100 countries around the world.
  https://mediacloud.org
- *Newsmap* is a very effective tool for tracking the current news topics in each country. *Newsmap* is an app that visually reflects the ever-changing landscape of *Google News*. The information is displayed in the form of treemaps that proportionally represent the most relevant topics from each country.
  https://newsmap-js.herokuapp.com

On the other hand, there are also paid tools that have been developed for digital marketing to analyze and manage the brands' work on social networks and that can also be used to track social media. Perhaps the best known are *Hootsuite*, *Buffer*, and *Fanpage Karma*.

From the standpoint of research in journalistic content analysis, we can consider some elements of interest that could be of value for academic teams:

- Focus: The platform allows access to informative content organized over time.
- Media coverage and different countries: It tracks relevant and diverse media from different countries to facilitate comparative studies.
- Accessibility: Free or registered or paid.
- Adaptability of enquiries: It allows the researcher to search for keywords and/or exact expressions configured –as unique or aggregated– with inclusion and exclusion criteria (Boolean operators, for example).
- Adaptability of visualizations: It allows the researcher to configure composition, relationship, comparison, and distribution graphs.
- Download of malleable corpus: You have the possibility to download a file with the textual contents and links related to the information.

Figure 1. *Newsmap* visualization interface

Table 1. Some comparable features

Factor	*Galileo*	*Media Cloud*	*Newsmap*
Focus	Social media networks/download of query data	Media on the web/download of query data	*Google News* (visualizations by country)
Coverage	9 countries	100+ countries	100+ countries
Accessibility	Upon request	Registered	Open access
Adaptability of queries	Boolean-based aggregate construction	Boolean-based aggregate construction	Predefined by each country's circumstances
Adaptability for visualizations	Multiple user-configurable visualizations	Those defined by the platform	Treemaps
Corpus download	Download in CSV or JSON file	Download in CSV file	Not available
Cost	No cost	No cost	No cost

2. *Galileo*: an option centered on social media journalism

To date, *Galileo* allows the integration of mined data from *Facebook*, *Instagram*, and *Twitter* APIs from 97 media outlets from nine countries (listed in the annex).
http://www.galileo-jrp.org

One could think of it as an IT infrastructure similar to those used by digital marketing services that seek to track brand accounts on social networks, capturing their content and their interactions. However, its main purpose is to aid in the search for patterns in the textual data of news media that circulate on social networks. *Galileo* is thus understood as a tool for extracting and filtering news from social networks.

In terms of information management, *Galileo* can be defined as a unified data management platform (UDMP). In other words, it is a centralized IT system that works with large amounts of structured data (frequencies and values) and unstructured data (texts) from diverse sources (*Facebook*, *Instagram*, and *Twitter*) to collect, integrate, manage, and visualize them.

2.1. IT architecture

One of the problems faced by platforms that collect a variety of information (textual corpora, values, and images, among others) is the flexibility required for their subsequent use. If the information is stored in fields structured as SQL databases, the end user will have the same restrictions that the programmer set when devising the classifications. Particularly, when working with textual data, excessive structuring will limit the researchers' possibilities for queries. *Galileo* aims to use a flexible NoSQL structure that uses *Elastic Stack*, a suite of open-source software products (*Apache 2.0* License) that allow the user to securely intake data from any source and format and then productively execute searches, analyses, and visualizations in real time or offline, as needed.
https://www.elastic.co

The *Galileo* crawler connects to the *CrowdTangle*[2] API to extract data from *Facebook* and *Instagram* or, in the case of *Twitter* posts, directly to this social network's API.

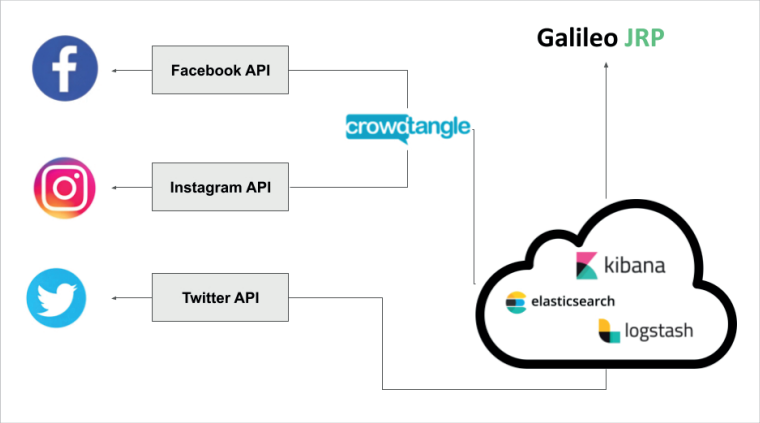

Figure 2. *Galileo* IT Architecture

A stack is a combination of tools, applications, and services that are used to create a web or mobile application. *Elastic Stack* (*ELK Stack*) comprises tools such as *Elasticsearch*, *Logstash*, and *Kibana* that offer particularly flexible functionalities for working with poorly structured or unstructured data such as press releases and social media posts. The malleability of *Elastic Stack* lies in having a NoSQL database (previously unstructured), which builds indexes from a set of data. An index contains mappings to various types of data, serving as a space for organizing information. Thus, *Galileo* is a kind of catch-all that stores the different types of data, but in organizational structure that does not confine them, precisely so that the users/researchers' consultations or queries are efficient, productive, and nearly unrestricted.

After collecting the data, *Galileo*'s greatest benefit is the ability to execute queries and provide data –which are not structured a priori– in a manageable way to perform content or discursive analysis. During these operations, the *Elastic Stack* suite offers three technically very efficient tools:

- *Elasticsearch*, or the core component of the *Elastic Stack*, is an open, distributed analytics and analysis engine for all types of data, including textual, numeric, geospatial, structured, and unstructured data, known for its simple *REST* APIs, distributed nature, speed, and scalability.
- *Logstash* is a free and open server-side data-processing pipeline that allows data to be taken in from a multitude of sources, transformed, and then stored and indexed in *Elasticsearch*.
- *Kibana* is a free and open front-end or interface application that is located on *Elastic Stack,* and it provides search capabilities on the data indexed in *Elasticsearch*, information visualization options, and the creation of dashboards with these visualizations.

Last but not least, it is possible to export the corpora of data built from the queries. *Elastic ELK* allows you to export data as CSV files that can then be easily read as spreadsheets or other tools that allow the development of automated analyses, such as *Jupyter*, *R Project*, and *spaCy* (**Guo** *et al.*, 2022). The use of this combination of components (ELK) as a tool suite has resulted in multiple adaptations. This is how we found applications for capacity control in massive venues (**Cecchet** *et al.*, 2020), process optimization in high-performance computing (**Underwood**, 2017), and organization of textual data from media (**Cárcamo-Ulloa** *et al.*, 2017).

2.2. Possibilities for research in journalism

Today, *Galileo* provides and makes more than 10 million news posts available to research communities. In this context, *Galileo* offers opportunities to carry out research with great volumes of textual data from media, identify patterns, and compare journalistic cultures (**Mellado** *et al.*, 2021a).

The data mined and available so far correspond to the posts made on *Facebook* (3,794,460), *Twitter* (6,474,089), and *Instagram* (369,960) between January 1, 2020, and December 31, 2021. In parallel, the crawler behind *Galileo* continues to mine data daily from the social media accounts of the media outlets integrated into the project (see list in Annex 1).

Table 2. Media followed by each country

Country	No. of media outlets	Language	Type of media				Average posts daily		
			Radio	TV	Press	Online	*Twitter*	*Facebook*	*Instagram*
Chile	10	Spanish	2	4	2	2	1,310.67	985.77	133.38
Spain	16	Spanish	4	4	4	4	1,838.53	1,014.89	55.66
Mexico	12	Spanish	3	2	5	2	1,369.27	842.13	38.45
United States	11	English	2	4	3	2	863.78	600.73	93.79
United Kingdom	15	English	4	4	4	3	1,417.09	698.25	42.36
Brazil	8	Portuguese	2	2	2	2	765.47	397.99	96.85
South Korea	8	Korean	2	2	2	2	333.71	223.61	5.79
Japan	9	Japanese	2	3	2	2	591.39	172.93	5.80
Germany	8	German	2	2	3	1	378.70	261.59	34.69

As we can see in Table 2, *Twitter* is the social network of choice for circulating daily news for the major media outlets of the nine countries participating in the project. However, the number of posts made on *Facebook* is substantial principally in Spanish-speaking countries. *Instagram* ranks third, far behind the other two social networks. Likewise, variations in the use of social networks between different countries can be observed, which reinforces the idea that journalistic cultures use different social network in different ways (**Mellado**; **Hermida**, 2021). For example, in Spain, the most intensive use is on *Facebook*, whereas in the other eight countries, the most frequent use continues to be on *Twitter*, which has traditionally been dominated by the political elite and opinion makers in each country. In communication studies, studies with *Twitter* have long been favored (**Arcila-Calderón**; **Barredo-Ibáñez**; **Castro**, 2017; **Hermida**, 2010). In addition to this network traditionally being used for news items of interest, this is probably due to the development of applications or plugins that have simplified accessing the *Twitter* API and downloading groups of texts associated with popular hashtags (#).

Table 3. Records available by country and social network

Country	*Twitter*		*Facebook*		*Instagram*		Total	%
	Posts	%	Posts	%	Posts	%		
Chile	956,786	15%	719,609	19%	97,371	27%	1,773,766	17%
Spain	1,342,124	21%	740,869	20%	40,633	11%	2,123,626	20%
Mexico	999,570	15%	614,753	16%	28,072	8%	1,642,395	15%
United States	630,556	10%	438,535	12%	68,470	19%	1,137,561	11%
United Kingdom	1,034,479	16%	509,722	13%	30,925	9%	1,575,126	15%
Brazil	558,794	9%	290,533	8%	70,701	19%	920,028	9%
South Korea	243,609	4%	163,235	4%	4,228	1%	411,072	4%
Japan	431,717	7%	126,242	3%	4,235	1%	562,194	5%
Germany	276,454	4%	190,962	5%	25,325	7%	492,741	5%
Total	6,474,089	100%	3,794,460	100%	369,960	100%	10,638,509	100%

In general terms, *Galileo* contains 10,638,509 media posts (see Table 3), with the ecosystem of Spain accounting for 20% of the dataset, followed by Chile with 17%, Mexico and the United Kingdom with 15% (each), the United States with 11%, and Brazil with 9%.

2.3. Operation

Galileo allows to create queries from keywords (See Figure 3). You can define the query [1] by (a) a set of keywords, (b) "exact expressions," or (c) combination with Boolean operators [AND, OR, NOT].

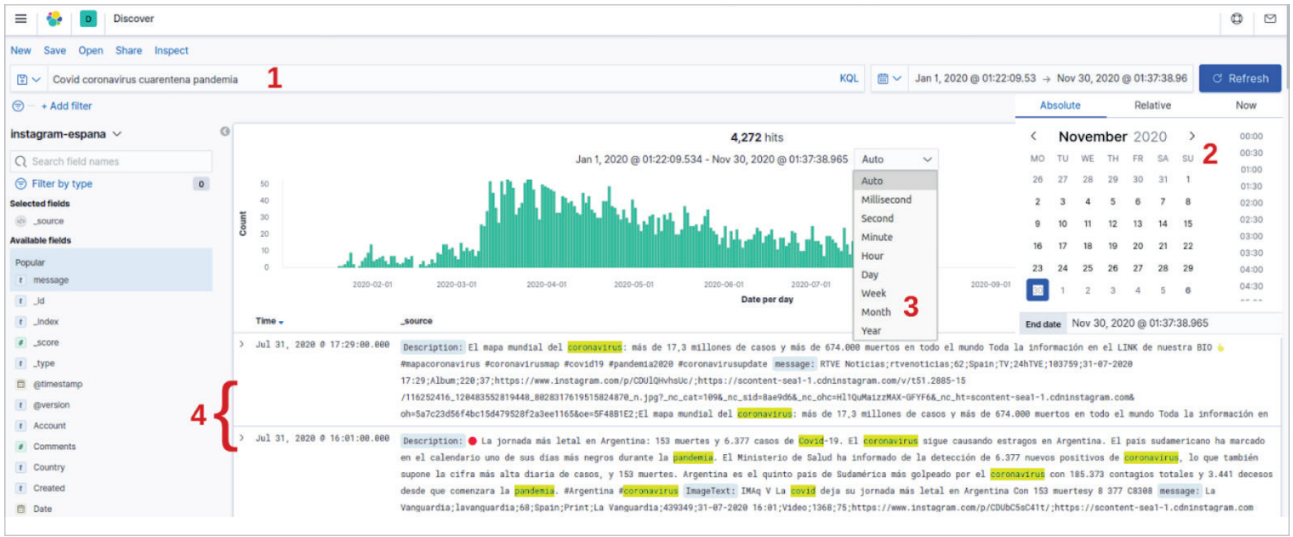

Figure 3. Keyword-based data search and filtering

La plataforma también permite determinar la fecha de entrada y salida de datos para la consulta [2] y permite elegir la periodicidad de las barras del histograma para visualizar la distribución en el tiempo de los datos consultados en forma diaria, semanal o mensual [3].

The platform also allows the user to determine the date of data input and output for the query [2] and to choose the frequency distribution of the histogram bars to visualize the distribution of the requested data over time on a daily, weekly, or monthly basis [3].

The data resulting from the query managed by the *Elasticsearch* engine can be organized and downloaded for further content analysis. To do this, it is necessary to organize the data, which otherwise maintain their formatting in JSON [4], before downloading it.

When sorting, saving, and downloading the data (see Figure 4), the user has several options and must choose an index (*Facebook*, *Twitter*, or *Instagram*) or a sub-index (each social network or each country) [1], selecting the fields they want to download (date, country, media, message, and url) [2].

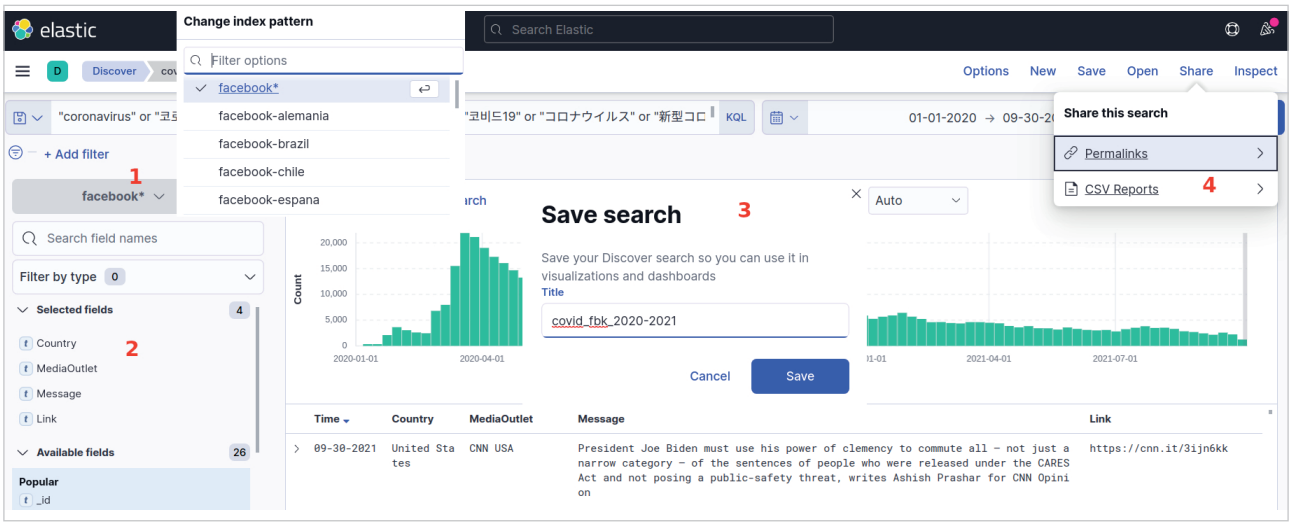

Figure 4. Sort, save, and download data in *Galileo*

The fields are different depending on the social network, but in all cases, the key fields that make up the post can be defined: name of the media outlet, texts, and URL to go to the message at its source.

Table 4. Post fields

Type of post	Downloadable fields
Facebook	Message, text of the headline (link) that leads to the news, name of the media outlet, date, country, URL for the original post, and followers at the time of posting
Instagram	Description (message), image text, name of media outlet, date, country, URL of the original post, and followers at the time of posting
Twitter	Message, name of media outlet, date, country, URL of the original post, and followers at the time of publication

Once the data output is ordered, as required, the results of the query must be saved [3] and the corpus must be downloaded in CSV format [4], in which each column is separated by a semicolon (;).

Note that the above figures display the default visualization of histograms. Figure 5 shows how the *Kibana* visualization builder allows for the easy configuration of different types of graphs. After defining the query period [1], the panel on the right [2] allows the user to choose the social network upon which the desired visualization will be run and [3] define the X and Y axes of the graph and its variables. An information popup that allows you to understand the values in detail, in this case for each country, is displayed when the cursor hovers over the graph [4]. If a multilingual query is performed [5], the graph will be redrawn [6] by simply clicking "refresh."

3. Three possibilities for the exploration of topics with *Galileo*

To demonstrate the potential of the *Galileo* platform, in this section, we will describe an analysis of the number of appearances of a set of keywords related to the term "climate change" in the press over the last 2 years[3].

3.1. The climate change scenario

In *Galileo,* a query about the presence of exact expressions linked to climate change was carried out in six languages:

- Spanish: "cambio climático" or "calentamiento global".
- English: "climate change" or "global warming".
- Portuguese: "aquecimento global" or "alterações climáticas".
- Japanese: "地球温暖化" or "気候変動".
- Korean: "지구 온난화" or "기후 변화".
- German: "globale Erwärmung" or "Klimawandel".

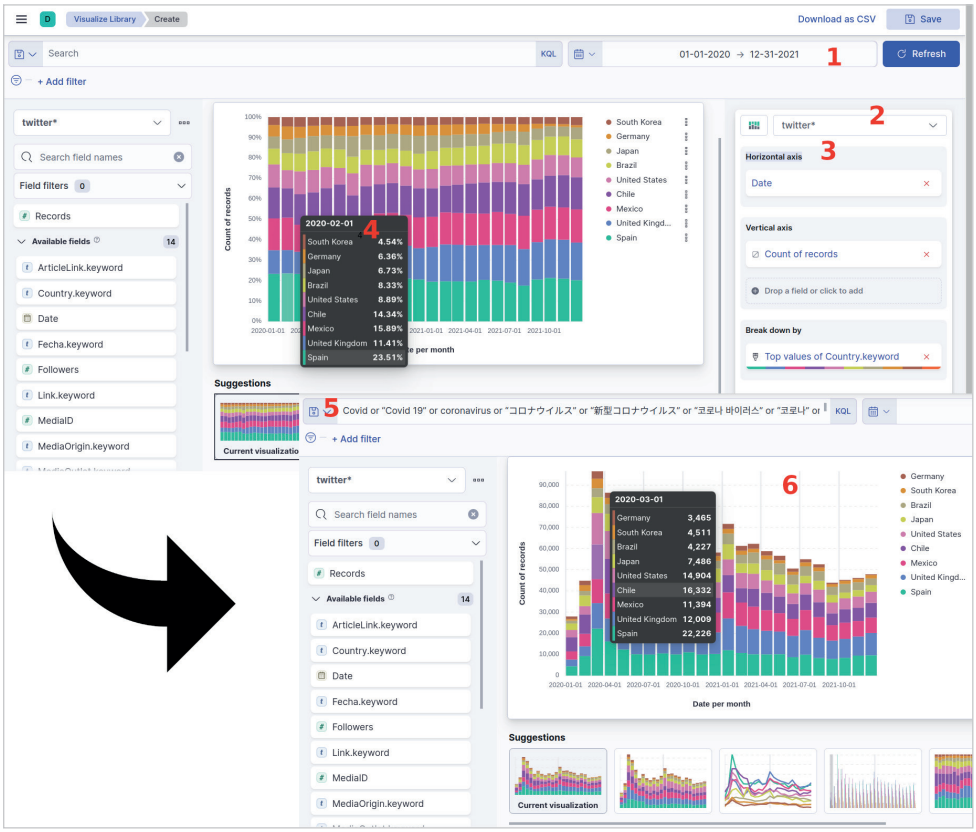

Figure 5. Visualizing data with *Kibana*

Table 5. Climate change data extraction

Countries	*Twitter*	*Facebook*	*Instagram*	Total
United Kingdom	5,145	2,035	596	7,776
United States	3,781	2,834	663	7,278
Chile	1,767	2,275	645	4,687
Spain	2,252	1,815	311	4,378
Mexico	1,523	1,536	127	3,186
Germany	961	870	616	2,447
Japan	770	308	51	1,129
Brazil	210	243	96	549
South Korea	12	13	5	30
Total	**16,421**	**11,929**	**3,110**	**31,460**

The result allows for the extraction of more than 31,460 posts for analysis. Anglo-Saxon countries mostly frequently refer to climate change, followed by Chile, Spain, and Mexico. For Germany, Japan, and South Korea, mentions drop considerably, and better results could certainly be found by refining queries country by country incorporating alternative keywords or linguistic variations that local specialists could suggest.

4. Conclusions

The enormous amount of textual data that circulates daily on social networks provides a great opportunity to carry out research in communication and journalism (**Arcila-Calderón**; **Barredo-Ibáñez**; **Castro**, 2017). Media outlets have moved their content to these platforms to maintain or improve the circulation of their news, opening up new research opportunities to understand information ecosystems. This type of research requires appropriate methodologies and tools to obtain and process such information (**Hamborg**; **Donnay**; **Gipp**, 2019). *Galileo* makes it possible to facilitate the processes of obtaining textual datasets from news media circulating on social networks. These corpora can be analyzed on the basis of automatic processing of language such as tokenization and lemmatization, among other techniques. Datasets are compatible with environments such as *Jupyter*, *R Project*, and *spaCy*, among others.

The experimental test of the platform was successful, and from January 2020 to December 2021, difficulties in accessing sources of information (social networks of media outlets) were not reported, supporting research done on news data science, such as the work of **Mellado** *et al*. (2021a; 2021b).

Researchers interested in using *Galileo* can follow news cycles in nine countries and examine relevant topics and the presence of specific entities. Obviously, a limited number of media were tracked, but we have taken care to focus on a relevant set for each country. *Galileo* can also be useful for NGOs or public services that monitor how issues or causes are presented in the mass media.

> *Galileo* provides and makes more than 10 million news posts available to research communities. Offers opportunities to carry out research with great volumes of textual data from media, identify patterns, and compare journalistic cultures

For content creators, *Galileo* can serve as a tool for understanding which narratives and/or narrative formats dominate a topic, allowing them to reverse engineer, in a way, the information development process or, as **Ojo** and **Heravi** (2018) proposed, analyze patterns and typologies of successful informative narratives.

Galileo's architecture is simple and robust. The user can connect directly to the *Twitter* API and indirectly to *Facebook* and *Instagram* through the *CrowdTangle* API. *Galileo* is offered as a tool that allows for the installation and configuration of a platform for tracking open, public social network accounts (it does not track personal profiles) and a resource for obtaining social media corpora from media outlets to study relevant phenomena such as polarization (**Levy**, 2021), fake news (**Cárcamo-Ulloa** *et al*., 2021), and confirmation bias (**Ling**, 2020).

5. Demonstrations and limitations

As a demonstration, the research team provides the username "invitado_epi" and the password "pass_epi", which will allow you to openly test *Galileo*[4], running queries and obtaining corpora according to your interests. Likewise, an invitation to request academic access and join the *Galileo* user community is available.

The creation of a user community will –going forward– extend *Galileo*'s capabilities to incorporate new social media information ecosystems and facilitate data mining for research. That is, new countries can be incorporated (beyond the nine countries currently tracked) and the number of media tracked in each country can be increased, generating potential databases for multiple research topics in communication and journalism.

Regarding its limitations, we can frankly say that (a) the tool still requires development to offer an automatic solution for requesting the incorporation of new media and other countries, (b) *Galileo* does not scrape content from the original media websites and, thereby, limits the working corpora to posts as a unit of analysis, and (c) although *Galileo* is able to contain the transcribed texts of the social media posts, it does not directly store the associated images. All of these limitations are challenges for future development.

6. Notes

1. For example, there are media groups such as *EMOL.com* from Chile that, from 2019 onwards, adopted the policy of publishing little news on social networks, favoring reader subscription on its website.

2. *CrowdTangle Team (2020). CrowdTangle. Facebook*, *Menlo Park*, *California*, *United States*.

3. We share the data from this extraction as open data for the academic community's use in communication sciences at: *https://github.com/luiscarcamo/cc_9c_2Y*

4. Access available at: *http://www.galileo-jrp.org*

You can also fill out the form available at: *https://forms.gle/VgJ1BAyuga1nTn4v5* to join the research community.

7. References

Arcila-Calderón, Carlos; **Barredo-Ibáñez, Daniel**; **Castro, Cosette** (2017). *Analítica y visualización de datos en Twitter*. Barcelona: Colección Comunicación. Editorial UOC. ISBN: 978 84 9116 960 4

Baden, Christian (2018). "Reconstructing frames from intertextual news discourse: A semantic network approach to news framing analysis". In: D'Angelo, Paul (ed.). *Doing news framing analysis II: Empirical and theoretical perspectives*. New York: Routledge, pp. 3-26. ISBN: 978 1 315642239

Baeza-Yates, Ricardo (2018). "Bias on the web". *Communications of the ACM*, v. 61, n. 6, pp. 54-61. *https://doi.org/10.1145/3209581*

Blei, David M.; **Ng, Andrew Y.**; **Jordan, Michael I.** (2003). "Latent Dirichlet allocation". *Journal of machine learning research*, v. 3, pp. 993-1022. *https://dl.acm.org/doi/10.5555/944919.944937*

Cárcamo-Ulloa, Luis; **Cárdenas-Neira, Camila**; **Sáez-Trumper, Diego**; **Toural-Bran, Carlos** (2021). "*Fake news* en Chile y España: ¿Cómo los medios nos hablan de noticias falsas?". *Journal of Iberian and Latin American research*, v. 26, n. 3, pp. 320-337. *https://doi.org/10.1080/13260219.2020.1909849*

Cárcamo-Ulloa, Luis; **Vernier, Matthieu**; **Scheihing-García, Eliana**; **Aravena, Matías**; **Pérez, Javier** (2017). "*Sophia* una herramienta para la construcción y análisis de casos noticiosos en la enseñanza del periodismo". *Nuevas ideas en informática educativa*, v. 13, pp. 667-672. ISBN: 978 956 19 1043 0
http://www.tise.cl/volumen13/TISE2017/96.pdf

Cárdenas-Neira, Camila (2016). "Representación online del movimiento estudiantil chileno: Reapropiación de noticias en *Facebook*". *Estudios filológicos* n. 58, pp. 25-49.
https://doi.org/10.4067/S0071-17132016000200002

Cardon, Dominique (2018). *Con qué sueñan los algoritmos: Nuestras vidas en los tiempos de los big data*. Madrid: Dado Ediciones. ISBN: 978 849 450 728 1

Cecchet, Emmanuel; **Acharya, Amrita**; **Molom-Ochir, Tergel**; **Trivedi, Amee**; **Shenoy, Prashant** (2020). "*WiFiMon*: a mobility analytics platform for building occupancy monitoring and contact tracing using wifi sensing: poster abstract". In: *Proceedings of the 18th conference on embedded networked sensor systems*, *SenSys'20*, pp. 792-793.
https://doi.org/10.1145/3384419.3430598

Grimmer, Justin; **Stewart, Brandon M.** (2013). "Text as data: The promise and pitfalls of automatic content analysis methods for political texts". *Political analysis*, v. 21, n. 3, pp. 267-297.
https://doi.org/10.1093/pan%2Fmps028

Guo, Lei; **Su, Chao**; **Paik, Sejin**; **Bhatia, Vibhu**; **Prasad-Akavoor, Vidya**; **Gao, Ge**; **Betke, Margrit**; **Wijaya, Derry** (2022). "Proposing an open-sourced tool for computational framing analysis of multilingual data". *Digital journalism*, first online.
https://doi.org/10.1080/21670811.2022.2031241

Hamborg, Felix; **Donnay, Karsten**; **Gipp, Bela** (2019). "Automated identification of media bias in news articles: an interdisciplinary literature review". *International journal on digital libraries*, n. 20, pp. 391-415.
https://doi.org/10.1007/s00799-018-0261-y

Hermida, Alfred (2010). "Twittering the news. The emergence of ambient journalism". *Journalism practice*, v. 4, n. 3, pp. 297-308.
https://doi.org/10.1080/17512781003640703

Jiang, Ye; **Song, Xingyi, Harrison, Jackie**; **Quegan, Shaun**; **Maynard, Diana** (2017). "Comparing attitudes to climate change in the media using sentiment analysis based on latent Dirichlet allocation". In: *Proceedings of the 2017 EMNLP workshop: Natural language processing meets journalism*, pp. 25-30.
https://doi.org/10.18653/v1/W17-4205

Levy, Ro'ee (2021). "Social media, news consumption, and polarization: evidence from a field experiment". *American economic review*, v. 111, n. 3, pp. 831-870.
https://doi.org/10.1257/aer.20191777

Lewis, Seth C.; **Zamith, Rodrigo**; **Hermida, Alfred** (2013). "Content analysis in an era of big data: a hybrid approach to computational and manual methods". *Journal of broadcasting & electronic media*, v. 57, n. 1, pp. 34-52.
https://doi.org/10.1080/08838151.2012.761702

Li, Yue; **Nair, Pratheeksha**; **Wen, Zhi**; **Chafi, Imane**; **Okhmatovskaia, Anya**; **Powell, Guido**; **Shen, Yannan**; **Buckeridge, David** (2020). "Global surveillance of Covid-19 by mining news media using a multi-source dynamic embedded topic model". In: *Proceedings of the 11th ACM international conference on bioinformatics, computational biology and health informatics*, *BCB'20*, n. 34.
https://doi.org/10.1145/3388440.3412418

Ling, Rich (2020) "Confirmation bias in the era of mobile news consumption: the social and psychological dimensions". *Digital journalism*, v. 8, n. 5, pp. 596-604.
https://doi.org/10.1080/21670811.2020.1766987

Matthes, Jörg; **Kohring, Mathias** (2008). "The content analysis of media frames: Toward improving reliability and validity". *Journal of communication*, v. 58, n. 2, pp. 258-279.
https://doi.org/10.1111/j.1460-2466.2008.00384.x

Mellado, Claudia; **Cárcamo-Ulloa, Luis**; **Alfaro, Amaranta**; **Inai, Darla**; **Isbej, José** (2021a). "Fuentes informativas en tiempos de Covid-19: Cómo los medios en Chile narraron la pandemia a través de sus redes sociales". *Profesional de la información*, v. 30. n. 4, e300421.
https://doi.org/10.3145/epi.2021.jul.21

Mellado, Claudia; **Hallin, Daniel**; **Cárcamo-Ulloa, Luis**; **Alfaro, Rodrigo**; **Jackson, Daniel**; **Humanes, María-Luisa**; **Márquez-Ramírez, Mireya**; **Mick, Jacques**; **Mothes, Cornelia**; **Lin, Christi-I-Hsuan**; **Lee, Misook**; **Alfaro, Amaranta**; **Isbej, José**; **Ramos, Andrés** (2021b). "Sourcing pandemic news: A cross-national computational analysis of mainstream media coverage of Covid-19 on *Facebook*, *Twitter*, and *Instagram*". *Digital journalism*, v. 9, n. 9, pp. 1261-1285.
https://doi.org/10.1080/21670811.2021.1942114

Mellado, Claudia; **Hermida, Alfred** (2021). "The promoter, celebrity, and joker roles in journalists' social media performance". *Social media + society*, v. 7, n. 1.
https://doi.org/10.1177/2056305121990643

Morstatter, Fred; **Wu, Liang**; **Yavanoglu, Uraz**; **Corman, Stephen R.**; **Liu, Huan** (2018). "Identifying framing bias in online news". *ACM Transactions on social computing*, v. 1, n. 2.
https://doi.org/10.1145/3204948

Newman, Nic; **Dutton, William H.**; **Blank, Grant** (2012). "Social media in the changing ecology of news: the fourth and fifth estates in Britain". *International journal of internet science*, v. 7, n. 1, pp. 6-22.
https://www.ijis.net/ijis7_1/ijis7_1_newman_et_al.pdf

Newman, Nic; **Fletcher, Richard**; **Schulz, Anne**; **Andi, Singe**; **Robertson, Craig T.**; **Nielsen, Rasmus-Kleis** (2021). *Digital news report 2021.* Reuters Institute for the Study of Journalism.
https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2021-06/Digital_News_Report_2021_FINAL.pdf

Ojo, Adegboyega; **Heravi, Bahareh** (2018). "Patterns in award winning data storytelling. Story types, enabling tools and competences". *Digital journalism*, v. 6, n. 6, pp. 693-718.
https://doi.org/10.1080/21670811.2017.1403291

Pariser, Ely (2011). *The filter bubble: How the new personalized web is changing what we read and how we think.* Penguin. ISBN: 978 0 143121237

Pereira, Moisés; **Cardeal-Pádua, Flavio-Luis**; **Machado-Pereira, Adriano-César**; **Silva, Giani-David**; **Benevenuto-de-Souza, Fabricio** (2015). "Multimodal sentiment analysis for automatic estimation of polarity tension of TV news in TV newscasts videos". In: *Proceedings of the 21st Brazilian symposium on multimedia and the web*, *WebMedia'15*, pp. 157-160.
https://doi.org/10.1145/2820426.2820461

Raimondo-Anselmino, Natalia; **Sambrana, Alejandro**; **Cardoso, Ana-Laura** (2017). "Medios tradicionales y redes sociales en internet: un análisis de los posteos compartidos por los diarios argentinos *Clarín* y *La Nación* en *Facebook* (2010-2015)". *Astrolabio*, n. 19, pp. 32-68.
https://revistas.unc.edu.ar/index.php/astrolabio/article/view/17787

Romanou, Angelika; **Smeros, Panayiotis**; **Castillo, Carlos**; **Aberer, Karl** (2020). "*Scilens* news platform: A system for real-time evaluation of news articles". In: *Proceedings of the VLDB endowment*, v. 13, n. 12, pp. 2969-2972.
https://doi.org/10.14778/3415478.3415521

Sáez-Trumper, Diego; **Castillo, Carlos**; **Lalmas, Mounia** (2013). "Social media news communities: gatekeeping, coverage, and statement bias". In: *Proceedings of the 22nd ACM international conference on information & knowledge management*, pp. 1679-1684.
https://doi.org/10.1145/2505515.2505623

Salazar, Diego (2019). *No hemos entendido nada: Qué ocurre cuando dejamos el futuro de la prensa a merced de un algoritmo*. Editorial Debate. ISBN: 978 84 17636258

Schmitz-Weiss, Amy; **De-Macedo-Higgins-Joyce, Vanessa**; **Saldaña, Magdalena**; **Calmon-Alves, Rosental** (2017). "Latin American investigative journalism education: Learning practices, learning gaps". *Journalism & mass communication educator*, v. 72, n. 3, pp. 334-348.
https://doi.org/10.1177/1077695817711611

Trilling, Damian; **Jonkman, Jeroen G. F.** (2018). "Scaling up content analysis". *Communication methods and measures*, v. 12, n. 2-3, pp. 158-174.
https://doi.org/10.1080/19312458.2018.1447655

Underwood, Richard (2017). "Building bridges: The system administration tools and techniques used to deploy bridges". In: *Proceedings of the practice and experience in advanced research computing 2017 on sustainability, success and impact*, *PEARC17*, article n. 5.
https://doi.org/10.1145/3093338.3093339

Van-Atteveldt, Wouter; **Peng, Tai-Quan** (2018). "When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science". *Communication methods and measures*, v. 12, n. 2-3, pp. 81-92.
https://doi.org/10.1080/19312458.2018.1458084

Vernier, Mathieu; **Cárcamo-Ulloa, Luis**; **Scheihing-García, Eliana** (2017). "Diagnóstico de la estrategia editorial de medios informativos chilenos en *Twitter* mediante un clasificador de noticias automatizado". *Revista austral de ciencias sociales*, n. 30, pp. 183-201.
https://doi.org/10.4206/rev.austral.cienc.soc.2016.n30-09

Watts, Duncan J. (2016). "Computational social science: Exciting progress and future challenges". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, *KDD'16*, p. 419. *https://doi.org/10.1145/2939672.2945366*

Watts, Duncan J. (2017). "Should social science be more solution-oriented?". *Nature human behaviour*, v. 1, artículo n. 15. *https://doi.org/10.1038/s41562-016-0015*

Watts, Duncan J.; **Rothschild, David M.**; **Mobius, Markus** (2021). "Measuring the news and its impact on democracy". In: Scheufele, Dietram (ed.). *Proceedings of the National Academy of Sciences*, v. 118, n. 15. *https://doi.org/10.1073/pnas.1912443118*

Zhang, Hao; **Boons, Frank**; **Batista-Navarro, Riza** (2019). "Whose story is it anyway? Automatic extraction of accounts from news articles". *Information processing & management*, v. 56, n. 5, pp. 1837-1848. *https://doi.org/10.1016/j.ipm.2019.02.012*

Zhou, Xinyi; **Zafarani, Reza**; **Shu, Kai**; **Liu, Huan** (2019). "*Fake news*: fundamental theories, detection strategies and challenges". In: *Proceedings of the 12th ACM international conference on web search and data mining*, *WSDM '19*, pp. 836-837. *https://doi.org/10.1145/3289600.3291382*

8. Annex

				Data mining		
Country (number of media outlets)	**Average**	**Type of media**	**Language**	***Twitter***	***Facebook***	***Instagram***
South Korea 8	*KBS*	TV	Korean	Yes	Yes	Yes
	JTBC	TV	Korean	Yes	Yes	Yes
	CBS South Korea	Radio	Korean	Yes	Yes	No
	TBS Radio	Radio	Korean	Yes	Yes	Yes
	Chosun	Print	Korean	Yes	Yes	Yes
	Hankyoreh	Print	Korean	Yes	Yes	No
	Dailian	Online	Korean	Yes	Yes	Yes
	OhMyNews	Online	Korean	Yes	Yes	Yes
Japan 9	*NHK*	TV	Japanese	Yes	Yes	Yes
	TV Asahi	TV	Japanese	Yes	Yes	Yes
	Nihon TV	TV	Japanese	Yes	Yes	Yes
	NHK radio	Radio	Japanese	Yes	Yes	Yes
	Bunka Housou	Radio	Japanese	Yes	Yes	No
	The Yomiuri Shinbum	Print	Japanese	Yes	Yes	Yes
	The Asahi Shinbum	Print	Japanese	Yes	Yes	Yes
	Huffpost Japan	Online	Japanese	Yes	Yes	Yes
	Buzzfeed Japan	Online	Japanese	Yes	Yes	Yes
Chile 10	*Mega*	TV	Spanish	Yes	Yes	Yes
	TVN	TV	Spanish	Yes	Yes	Yes
	Canal 13	TV	Spanish	Yes	Yes	Yes
	CNN Chile	TV	Spanish	Yes	Yes	Yes
	Biobio.cl	Radio	Spanish	Yes	Yes	Yes
	Emol.com	Online	Spanish	Yes	Yes	Yes
	Elmostrador.cl	Online	Spanish	Yes	Yes	Yes
	Cooperativa	Radio	Spanish	Yes	Yes	Yes
	La Tercera	Print	Spanish	Yes	Yes	Yes
	Las Últimas Noticias	Print	Spanish	Yes	Yes	Yes
México 12	*Televisa*	TV	Spanish	Yes	Yes	Yes
	Canal Once	TV	Spanish	Yes	Yes	No
	Radio Fórmula	Radio	Spanish	Yes	Yes	Yes
	Aristegui Noticias	Radio	Spanish	Yes	Yes	Yes
	Instituto Mexicano de la Radio	Radio	Spanish	Yes	Yes	No
	La Jornada	Print	Spanish	Yes	Yes	Yes
	Reforma	Print	Spanish	Yes	Yes	Yes
	El Financiero	Print	Spanish	Yes	Yes	Yes
	La Prensa	Print	Spanish	Yes	Yes	Yes
	Animal Político	Online	Spanish	Yes	Yes	Yes
	El Universal Online México	Print	Spanish	Yes	Yes	Yes
	UnoTV	Online	Spanish	Yes	Yes	Yes

Brazil 8	*Record*	TV	Portuguese	Yes	Yes	Yes
	CBN	Radio	Portuguese	Yes	Yes	Yes
	Band	Radio	Portuguese	Yes	Yes	Yes
	Folha de S. Paulo	Print	Portuguese	Yes	Yes	Yes
	O Globo	TV	Portuguese	Yes	Yes	Yes
	O Estado de S. Paulo	Print	Portuguese	Yes	Yes	Yes
	R7	Online	Portuguese	Yes	Yes	Yes
	G1	Online	Portuguese	Yes	Yes	No
United States 11	*NBC*	TV	English	Yes	Yes	Yes
	CNN USA	TV	English	Yes	Yes	Yes
	KABC	TV	English	Yes	Yes	Yes
	NPR	Radio	English	Yes	Yes	Yes
	CBS US	Radio	English	Yes	Yes	Yes
	New York Times	Print	English	Yes	Yes	Yes
	USA Today	Print	English	Yes	Yes	Yes
	Los Angeles Times	Print	English	Yes	Yes	Yes
	Buzzfeed US	Online	English	Yes	Yes	Yes
	Fox US	TV	English	Yes	Yes	Yes
	Huffington Post US	Online	English	Yes	Yes	Yes
Spain 16	*Telecinco*	TV	Spanish	Yes	Yes	Yes
	Antena 3	TV	Spanish	Yes	Yes	Yes
	La Sexta	TV	Spanish	Yes	Yes	Yes
	TVE	TV	Spanish	Yes	Yes	Yes
	SER	Radio	Spanish	Yes	Yes	Yes
	COPE	Radio	Spanish	Yes	Yes	Yes
	Onda Cero	Radio	Spanish	Yes	Yes	Yes
	Radio Nacional de España	Radio	Spanish	Yes	Yes	Yes
	El Pais	Print	Spanish	Yes	Yes	Yes
	La Vanguardia	Print	Spanish	Yes	Yes	Yes
	El Mundo	Print	Spanish	Yes	Yes	Yes
	ABC Spain	Print	Spanish	Yes	Yes	Yes
	El Confidencial	Online	Spanish	Yes	Yes	Yes
	OK Diario	Online	Spanish	Yes	Yes	Yes
	Eldiario.es	Online	Spanish	Yes	Yes	Yes
	Huffpost ES	Online	Spanish	Yes	Yes	Yes
Germany 8	*ARD*	TV	German	Yes	Yes	Yes
	RTL	TV	German	Yes	Yes	Yes
	Deutschland-funk	Radio	German	Yes	Yes	Yes
	Klasssík Radio	Radio	German	Yes	Yes	Yes
	BILD	Print	German	Yes	Yes	Yes
	Frankfurter Allgemeine Zeitung	Print	German	Yes	Yes	Yes
	Süddeutsche Zeitung	Print	German	Yes	Yes	Yes
	Spiegel Online	Print	German	Yes	Yes	Yes
United Kingdom 15	*BBC News*	TV	English	Yes	Yes	Yes
	Channel 4 News	TV	English	Yes	Yes	Yes
	Sky News	TV	English	Yes	Yes	Yes
	ITV News	TV	English	Yes	Yes	Yes
	BBC Radio 4	Radio	English	Yes	Yes	No
	BBC Radio 2	Radio	English	Yes	Yes	Yes
	TalkSport	Radio	English	Yes	Yes	Yes
	ClasSíc FM	Radio	English	Yes	Yes	Yes
	The Daily Telegraph	Print	English	Yes	Yes	Yes
	The Guardian	Print	English	Yes	Yes	Yes
	The Daily Mirror	Print	English	Yes	Yes	Yes
	The Sun	Print	English	Yes	Yes	Yes
	Mail online	Online	English	Yes	Yes	Yes
	Huffpost UK	Online	English	Yes	Yes	Yes
	Buzzfeed UK	Online	English	Yes	Yes	Yes