TRI-POL: selecting the lists of news media domains and subdomains for each country

Oriol J. Bosch
London School of Economics and Political Science

June 8, 2021

For the TRI-POL project we tracked individuals' online behaviours in order to create variables measuring individual's media consumption. We wanted to create variables measuring more general behaviours as 1) the number of visits and time spent visiting news media outlets; as well as more specific behaviours as 2) the time spent visiting specific content within those outlets, for instance political articles. You can find a list and description of the metered data variables created for this project in the *Documentation and data archive*.

Most past research has only used domain information (e.g. *theguardian.com*) to compute these types of variables. Nonetheless, we believe that if wanting to know whether individuals are exposed to more specific content, like political information, only looking at the domain level might produce variables with low validity. For instance, a participant visiting *theguardian.com* can read articles about politics but also about sports, culture or society among others topics. If wanting to measure the effect of people reading about politics, not excluding those articles about other topics might confound the results.

Therefore, to create these variables, there was the need to define which news media outlets should be tracked and, within those news media outlets, what URLs should be considered as political or not. Hence, we had to define two things:

- The list of domains to track (e.g. *theguardian.com, thetimes.co.uk, bbc.co.uk*, etc).
- The list of subdomains to track within those domains (e.g.*theguardian.com/opinion/, theguardian.com/politics/, theguardian.com/world/*, etc)

There is no consensus in the literature on how to find and create lists of news media outlets, nor how to define what subdomains should be considered as political or not. Nonetheless, these decisions will determine how valid our measures are and how comparable they are across countries. In this document, therefore, we explain the strategy that we followed in order to identify and select the list of domains and subdomains used to compute our metered data variables.

Identifying the list of all the news outlets with a web page in each country

The first step was to identify a list of all the news media outlets with a web page, in each of the countries. Hence, this means purely digital outlets (e.g. *buzzfeed.com*), traditional ones with a web page (e.g. *theguardian.com*) and also TVs and radios with a web page where they upload written news (e.g. *bbc.com.uk*). This list needed to be exhaustive, covering not only national outlets but also regional ones. For instance, in Spain we identified a list of 776 outlets.

The main challenge faced in this step was that there is no consensus on how to find or elaborate these lists. Some researchers have used the top 500 list of most popular news outlets created by Alexa (Cardenal, Aguilar-Paredes, Galais, & Pérez-Montoro, 2019). This list, nonetheless, is no longer available since September 2021. Other researchers have used official open-access information (e.g. Germany, Scharkow, Mangold, Stier, and Breuer 2020). However, for most countries this is not an option. And some others have used lists created by academics (Guess, Barberá, Munzert, & Yang, 2021), which again are not available for every country.

Considering all this, we asked our national teams to find a comprehensive list for their country, either from official sources, auditing companies or other trusted sources (e.g. created by academics or media professionals). Knowing the limitations, nonetheless, we did not expect the lists to be complete, which might have been impossible, but wanted them to be 1) big enough to give us confidence that we were covering both mainstream/fringe and national/regional media; and 2) coherent in the sense that it includes some of the most important media outlets in the country. Because some lists might not be perfect, we also instructed the national teams to manually edit the lists to include outlets which might have not been included, or exclude those which were not of interest for us (e.g. sport outlets). For instance, in Spain the list that we used as a base did not include TVs and radios. We manually included them because individual's also use these web pages to read news. The complete lists and the explanation of where they come from can be found in the *Documentation and data archive*.

Choosing which news media outlets to use

There is no common practice in the literature about which news media outlets to use to create metered data variables. Some have used all the news media outlets (Mangold, Stier, Breuer, & Scharkow, 2021), whether others have only used a list of the most visited (e.g. top 42, Cardenal et al. 2019; top 50, Bach et al. 2019). Using information from all news media outlets might not be efficient, since most of them might not even be visited by participants. For instance, manually choosing the subdomains to include for hundred of domains, for each country (e.g. +700 domains in Spain), would not be feasible. Past research has found, indeed, that the 42 most visited news media outlets accounted for 99.85 per cent of all reported visits to online news outlets in a sample of individuals from the very same panel that we are using (Cardenal et al., 2019). However, if we exclude some domains, which ones should be excluded? Excluding too

many news media outlets might systematically bias the variables created. For instance, if most regional news media outlets are excluded, and there is a specific profile that only consumes regional news media outlets, we would not observe their behaviour.

Considering this, we decided to compute the same variables using different number of outlets:

- All domains
- 300 most visited domains
- 100 most visited domains
- 50 most visited domains
- 10 most visited domains

Ranking the news media outlets from most visited to least visited

To create different lists of news media outlets (i.e. all, top 300, top 100, top 50, top 10), we needed to rank the identified news media outlets from most visited to least visited. However, as said before, Alexa does not provide with a list of the most visited outlets anymore. Nonetheless, we found in Tranco a solid and high-quality alternative. Tranco is a research-oriented top sites ranking hardened against manipulation, which is a more transparent and open-source alternative to sites like Alexa. Essentially, they combine and curate three of the most popular rankings (Alexa, Umbrella and Majestic), to create a more stable and reliable list of the most visited web pages worldwide. The main drawback is that it does not allow to differentiate by categories nor countries. Nonetheless, it provides with a ranking of the 1 million most visited/popular web pages in the world, ranked from most visited to least visited. This is enough to create a ranking of the most visited digital media outlets in each of our countries.

How did we do that? We combined our lists of the news media outlets in each of the countries with the list of the 1 million most visited web pages created by Tranco. For each ountry, those web pages that were in both lists, were kept, while the others were excluded. In Spain, for example, we identified around 300 news media outlets in the Tranco list. Next, using information from the Tranco list, we ranked the news media outlets from most visited to least visited. After creating the ranked lists of news media outlets for each country, we checked them to see if they were coherent. Hence, we checked whether there were mainstream/fringe and national/regional outlets included, as well as whether the ranking resembled other existing rankings (e.g. comScore, national auditing). All the lists passed this check.

The Stata scripts used to combine both data sources and create the ranked list can be found in the *Documentation and data archive*. In addition, the ranked lists can also be found there.

Creating the different lists of most visited news media outlets

Next, we wanted to create the different lists. Essentially, we only kept a limited number of domains, based on their ranking. How did we create these list? Using the previously

created ranked list of news media outlets, we selected only the most visited domains according to Tranco. In the *Documentation and data archive* you can check the Stata scripts that we used. The different lists for each country can also be found there.

Selecting the subdomains

As said before, we needed to define not only the domains (e.g. *theguardian.com*) but also the subdomains (e.g.*theguardian.com/opinion/*). Although this could be done for the complete lists for each country, the amount of effort and time that it would require was considered too big (e.g. +700 hundred domains only in Spain). Therefore, we decided to collect subdomain information only for the 50 most visited news media outlets in each country. Past research has found, indeed, that the 42 most visited news media outlets accounted for 99.85 per cent of all reported visits to online news outlets in a sample of individuals from the very same panel that we are using (Cardenal et al., 2019).

Using the top 50 most visited news media outlets for each country, we needed to define which subdomains to track within each of them. More specifically, we wanted to know whether participants visited 1) political articles and, within this broad category, 2) national politics articles, 3) regional politics articles, 4) international politics articles and 5) opinion articles. Hence, we needed to define which URLs to consider as part of these categories. Different approaches could be followed to do so. The URLs could be manually checked and codified as political or not, or a machine learning algorithm could be trained to automatically identify the content of an article as political or not. Nonetheless, for this project we do not have access to the full URLs, but only to domain and subdomain information. Hence, we needed to develop a strategy to identify the content of an article only using subdomain information. Luckily, most news media outlets use different subdomains to identify the content of their articles e.g. */politics/* for political news or */national/* for news about national politics. Although this is obviously not an error free approach, we believe it allows for a better accuracy than considering all URLs as political.

The main challenge to develop this approach, nonetheless, is that not all websites organize their URLs in the same way i.e. not all political subdomains are labeled as */politics/*. It is not, therefore, a straightforward task. Following a set of rules that we designed beforehand, every national team manually selected the URLs falling in each of the previously presented categories.

Political news:

- If the website has a specific subdomain */politics/* this must be included.
- If the website has a specific subdomain */national* (e.g. Portugal)/ this must be included.
- If the website has a specific subdomain */international/* or something like this (e.g. */world/*) this must be included.
- If the website has a specific subdomain */opinion/* this must be included.
- Some websites might have regional subsections e.g. */cataluna/*. Or even city subsections */London/*. This must be included too.

- There are some websites that do not present usable subdomains. For instance, some present as a subdomain the date of the article e.g. */12-05-21/*. In this case we cannot specify any subdomain and all should be considered political.
- Sometimes the subdomains can be defined by looking at the categories (e.g. national, international, sports, culture) of the web page, or at the menu. However, the true subdomains appear on the URLs. Hence, some websites might have the category "international", which brings you to a new page with all the news about international affairs. However, when you enter the article, you see that the real subdomain is */politics/*. This means that */international/* is not actually a subdomain, but a tag: they put together international news in a section, but their true subdomain is */politics/*.

National news:

- If the website has a specific subdomain */national* (e.g. Portugal)/ this must be included.
- If the website also has a subdomain */politics/*, check which news they categorize with this subdomain. If they are used only for national news, consider them as national. Besides, if the outlet only uses the subdomain */politics/* but it is almost entirely focused on national news, also consider it as national.
- There are some websites that do not present usable subdomains. For instance, some present as a subdomain the date of the article e.g. */12-05-21/*. We cannot differentiate between political news.

Regional news:

- Some websites might have regional subsections e.g. */cataluna/*. Or even city subsections */London/*. This must be included too.
- If the website also has a subdomain */politics/*, check which news they categorize with this subdomain. If they are used only for regional news, consider them as regional. Besides, if the outlet only uses the subdomain */politics/* but it is almost entirely focused on regional news, also consider it as regional.
- There are some websites that do not present usable subdomains. For instance, some present as a subdomain the date of the article e.g. */12-05-21/*. We cannot differentiate between political news.

International news:

- If the website has a specific subdomain */international/* or something like this (e.g. */world/*) this must be included.
- If the website also has a subdomain */politics/*, check which news they categorize with this subdomain. If they are used only for international news, consider them as international. Besides, if the outlet only uses the subdomain */politics/* but it is almost entirely focused on international news, also consider it as international.

- There are some websites that do not present usable subdomains. For instance, some present as a subdomain the date of the article e.g. */12-05-21/*. We cannot differentiate between political news.

Opinion pieces:

- If the website has a specific subdomain */opinion/* this must be included.
- If the website also has a subdomain */politics/*, check which news they categorize with this subdomain. If they are used only for opinion pieces, consider them as opinion. Besides, if the outlet only uses the subdomain */politics/* but it is almost entirely focused on opinion pieces, also consider it as opinion.
- There are some websites that do not present usable subdomains. For instance, some present as a subdomain the date of the article e.g. */12-05-21/*. We cannot differentiate between political news.

Using these guidelines, the national teams selected the subdomains to be considered in each of the categories. Figure 1 shows an excerpt of the list created for Spain. You can check the complete lists for all countries n the *Documentation and data archive*.

Ranking	Domain level	Political subdomain	National	Regional	International	Opinion
1	elpais.com	https://elpais.com/internacional/ https://elpais.com/opinion/ https://elpais.com/espana/ https://elpais.com/noticias/andalucia/ https://elpais.com/noticias/pais-vasco/ https://elpais.com/noticias/europa/ https://elpais.com/noticias/estados-unidos/ https://elpais.com/noticias/mexico/ https://elpais.com/noticias/latinoamerica/ https://elpais.com/noticias/oriente-proximo/ https://elpais.com/noticias/asia/ https://elpais.com/noticias/africa/	https://elpais.com/espana/	https://elpais.com/noticias/andalucia/ https://elpais.com/noticias/pais-vasco/	https://elpais.com/internacional/ https://elpais.com/noticias/europa/ https://elpais.com/noticias/estados-unidos/ https://elpais.com/noticias/mexico/ https://elpais.com/noticias/latinoamerica/ https://elpais.com/noticias/oriente-proximo/ https://elpais.com/noticias/asia/ https://elpais.com/noticias/africa/	https://elpais.com/opinion/
2	elmundo.es	https://www.elmundo.es/espana/ https://www.elmundo.es/madrid/ https://www.elmundo.es/andalucia/ https://www.elmundo.es/baleares/ https://www.elmundo.es/cataluna/ https://www.elmundo.es/comunidad-valenciana/ https://www.elmundo.es/pais-vasco/ https://www.elmundo.es/opinion/	https://www.elmundo.es/espana/	https://www.elmundo.es/madrid/ https://www.elmundo.es/andalucia/ https://www.elmundo.es/baleares/ https://www.elmundo.es/cataluna/ https://www.elmundo.es/comunidad-valenciana/ https://www.elmundo.es/pais-vasco/	https://www.elmundo.es/internacional/	https://www.elmundo.es/opinion/

Figure 1: Excerpt of the list of subdomains for Spain

References

Bach, R. L., Kern, C., Amaya, A., Keusch, F., Kreuter, F., Hecht, J., & Heinemann, J. (2019, oct). Predicting Voting Behavior Using Digital Trace Data. *Social Science Computer Review*, 089443931988289. Retrieved from https://doi.org/10.1177%2F0894439319882896 doi: 10.1177/0894439319882896

Cardenal, A. S., Aguilar-Paredes, C., Galais, C., & Pérez-Montoro, M. (2019, jul). Digital Technologies and Selective Exposure: How Choice and Filter Bubbles Shape News Media Exposure. *The International Journal of Press/Politics*, *24*(4), 465–486. Retrieved from https://doi.org/10.1177%2F1940161219862988 doi: 10.1177/1940161219862988

Guess, A. M., Barberá, P., Munzert, S., & Yang, J. (2021, mar). The consequences of online partisan media. *Proceedings of the National Academy of Sciences*, *118*(14), e2013464118. Retrieved from https://doi.org/10.1073%2Fpnas.2013464118 doi: 10.1073/pnas.2013464118

Mangold, F., Stier, S., Breuer, J., & Scharkow, M. (2021, jan). The overstated generational gap in online news use? A consolidated infrastructural perspective. *New Media & Society*, 146144482198997. Retrieved from https://doi.org/10.1177%2F1461444821989972 doi: 10.1177/1461444821989972

Scharkow, M., Mangold, F., Stier, S., & Breuer, J. (2020, jan). How social network sites and other online intermediaries increase exposure to news. *Proceedings of the National Academy of Sciences*, *117*(6), 2761–2763. Retrieved from https://doi.org/10.1073%2Fpnas.1918279117 doi: 10.1073/pnas.1918279117