THE QUALITY OF GOVERNMENT INSTITUTE

UNIVERSITY OF GOTHENBURG

The hitchhiker's guide to web-mediated text

Method handbook for quantification of online linguistic data in a country-specific context

Official research report,
Linguistic Explorations of Societies
(Work Package 1)

2022

Working paper series 2022:1

Jonas Andersson Schwarz

THE QUALITY OF
GOVERNMENT INSTITUTE

GÖTEBORGS
UNIVERSITET

The hitchhiker's guide to web-mediated text

Method handbook for quantification of online linguistic data in a country-specific context

Official research report, Linguistic Explorations of Societies (Work Package 1)

Jonas Andersson Schwarz

WORKING PAPER SERIES 2022:1

QoG THE QUALITY OF GOVERNMENT INSTITUTE
Department of Political Science
University of Gothenburg
Box 711, SE 405 30 GÖTEBORG
March 2022
ISSN 1653-8919

The hitchhiker's guide to web-mediated text

Method handbook for quantification of online linguistic data in a country-specific context

Official research report, ***Linguistic Explorations of Societies*** *(Work Package 1)*

Jonas Andersson Schwarz (Södertörn University, Sweden)

This report was written 2019–2021 as the main research report for Work Package 1 in which I was the Principal Investigator, as part of the cross-disciplinary research project ***Linguistic Explorations of Societies*** *(LES) funded by the Swedish Research Council (Vetenskapsrådet).*

I would want to express my gratitude to the many students, scholars, and experts who have in different ways contributed to the report, by kindly offering constructive feedback and insights:

Stefan Dahlberg (Mid Sweden University) and **Sofia Axelsson** (Department of Political Science, University of Gothenburg) who have been great convenors and administrators of the LES project; **Valeria Caras** (political-historical overview of the 20 countries); **Jesse Salazar** (data entry, analysis, proof-reading, feedback); **Fredrik Olsson**'s astute observations about the specificities of the supply side of web-mediated language data, including his useful checklist for data providers (see, e.g., Table 2); **Amaru Cuba Gyllensten** (classifier on the Swedish news text data); **Johan Hammarlund** (Kairos Future) who gave feedback on early draft; **Isabel Löfgren** (Brazil); **Paola Sartoretto** (Brazil); **Amaraizu Iheanyi Genius** (Nigeria); **Jessica Gustafsson** (Kenya); **Ling-Yi Huang** (Hong Kong); **Liisa Sömersalu** (Estonia); **Philipp Seuferling** (Germany).

SECTIONS

1. Introduction

1.1 Introduction

The purpose of this report is to provide a guide to social-science scholars, researchers, and students to the contemporary ecosystem of editorially generated and user-generated text on the World Wide Web – with a particular eye to the **transnational nature of the contemporary global media ecosystem** and conditions for **country-comparative analyses**. We will explore both data providers and the state of play of online media in various countries. The guide primarily concerns **issues of validity and reliability of internet-mediated text data** for computational linguistics and quantitative social sciences.

Since the scale of the internet has exploded in recent decades, the diversity in terms of languages and regional origin is considerable, as is the range of sources. Hence, this report is intended to sift through many of the possible challenges and biases – practical and theoretical – that are found on the **supply side of internet-mediated text**. In addition to this, the report will detail various factors that are important when **contextualising various linguistic and demographic specificities**. By using a sample of 20 manually selected countries, many such factors are addressed.

One of our intentions with the report is to **explore the possibilities** for an analyst to make **large-scale linguistic analysis**, involving **several different languages**, without necessarily being proficient in all the languages in question. This is a somewhat controversial, but analytically interesting and – in the context of a world society where the need for rapid assessments of transnational flows of digitally mediated communication – hugely important question. It is worth mentioning, from the outset, that we strongly advise any researcher who endeavours to explore the field to have participants who, to as large an extent as possible, have contextual knowledge about the languages/regions in question. With very large, computer-based studies of linguistic distributions in various languages and countries, this is however not always feasible – especially when approaching text corpuses as statistical entities, as a lot of contemporary, machine-learning oriented data linguistics does, and especially when time is of the issue, like with many of today's challenges to rapidly counter global dissemination of disinformation.

Hence, various important provisos and heuristics are necessary, to guide macroscopic researchers so that they manage to avoid some of the possible pitfalls that are sometimes observable from a social-science point of view. By critically considering these various factors, **this report will serve as a handbook** that can help interrogative researchers by addressing important stumbling blocks that are likely to be decisive for ensuring tolerable reliability and validity.

After an initial **statement of aims and challenges**, we briefly introduce an **epistemological and methodological discussion** about key research values such as reliability, validity, and representativity. Thereafter, the report delves into a more concrete account of the ecosystem of so-called **online text providers** (practically, all of these are private enterprises, which poses a few challenges for independent research purposes) and, also, the more specific **conditions for different countries and languages**, using 20 selected countries as case studies. Since one of the key purposes of the report is to provide a toolkit for comparative linguistic and geographic analysis, manifest indicators of factors like language communities, quality of government, press freedom etc. are as important as the metrics of relative popularity for different editorial and social media sources in each country.

Our empirical overview aims at concretely exploring what news titles exist on the open Web for each selected country. First, we try to assess the **relative provenience/validity for these titles**. Do they seem to provide conventional text-based news reporting, are they propaganda sites, or, alternatively, pure-play broadcasting/streaming sites with very little text content? One needs to filter away those sources that have low face validity, in order to try and get as fair and typical a selection as possible of actual editorial news sources. Second, we will try and estimate, through using Web traffic data (top-domain URLs), **a measure of relative comparability of popularity** for these various news sources.

Our selected countries are the following: Brazil, Egypt, Estonia, France, Germany, Great Britain, Hong Kong, Hungary, Indonesia, Italy, Kenya, Malaysia, Mexico, Nigeria, Poland, Russia, South Africa, Spain, Sweden, and USA. We have selected these countries, based on various factors: an intention to cover **a broad and diverse sample of countries, combined with a specific interest in countries that have either seen significant political shifts or domestic polarization in recent years** (e.g., Brazil, Egypt, Great Britain, Hong Kong, Hungary, Poland, Russia, USA), and contextual familiarity on behalf of us as researchers and authors of this report. The report also serves as a historical document, surveying the last decade's media development in these countries.

1.2 Research challenges for data-driven linguistic explorations of societies: access, provenance, validity, representativity

As a multidisciplinary collaboration between social scientists, computer scientists, and computational linguists, the Linguistic Explorations of Societies (LES) project draws on recent developments in natural language processing to meet the challenges facing the changing landscape of comparative survey research. Using language technology in conjunction with online text data from many languages

and countries, the aim has been to address issues of survey item comparability and measurement equivalence in cross-cultural surveys, with a particular focus on survey translations.

One key factor for this comparative approach to be valid is that the online text data from various countries is of the apposite kind. If such "found" online text is to be said to be useful for comparative social science, it ought to be minimally **reliable** and **accessible**. Large quantities of aggregated Web-mediated text are available from commercial providers, but it is not entirely clear that such collections are suitable for the rather more stringent requirements for **validity** demanded by social science, compared to, e.g., marketing. At the same time, being a highly commercialized environment, surprisingly few large-scale, multilingual collections of Web-mediated text exist, besides these commercial providers, making it a natural choice for social scientists to nevertheless investigate the potential usefulness of such resources of commercially scraped online text. While non-profit initiatives exist, and national libraries in various countries routinely scrape the Web, the purpose of this report is to explore the usefulness of commercially harvested online text, and the requisites that ought to be fulfilled in order to make such data suitable to serious academic research.

In the LES project, the very specific samples used – plaintext survey responses, alongside editorial and social textual discourse on the Web – shall, consequently, not be mistaken to be directly representative of entire languages or demographics. At the same time, they are intended to be very indicative samples of **intralinguistic properties of languages**. For the integrity and reliability of the project, it is important that adequate choices are made, so as to balance these two aspects of **linguistic representativity** and requisite **source-provenance validity**. An important step towards more equitable such considerations is to improve the possibilities for properly **contextualizing** the source material in question: Web-mediated online text.

In order to be able to contextualize and assess collections of online-mediated text for quality and validity, one would conventionally be required to speak the language in question, and ideally have some knowledge about the particular demographic, political, and historical conditions in the country or region in question. However, in multilingual and transnational research projects, one cannot always assume that the research team would have the luxury of having a set of researchers, covering all possible languages that one would like to make large-scale data-driven inferences from. The purpose of this report is, therefore, to explore what possible mistakes to avoid and challenges to heed, when managing online text data from many languages and countries for data-driven, statistical language analysis – with the specific intent to explore a set of selected cases, revealing some useful considerations, and even practical heuristics and approaches that might help the researcher to ask more inquisitive questions, and not settle with

what is (commercially) available, but instead trying to address potential shortfalls on the current supply side of Web-mediated online text.

As representativity and replicability are critical criteria for any research reliant on quantitative methodology, the limitations addressed below naturally raise cause for concern. It is often hard to control whether the data content of a given language and/or country from one media type is largely representative of the media in that very language and/or country, and whether data from the source types are comparable across different languages and countries. Although we can access information about the top URLs from which the models sample the largest amounts of web documents, it is difficult to assess the representativity of such platforms for several reasons. First, an examination of this kind requires intimate knowledge of internet usage in each country. For instance, a non-Swahili speaking person without familiar knowledge of Kenyan online media would have a difficult time to assess the content validity of Swahili web documents from Kenya. Second, even if we had access to this knowledge and were able to assess content validity, there is no straightforward method for assessing whether existing source URLs or specific web-documents within a given source URL are representative of the internet population of a given country. Disaggregated data on platform usage – particularly where demographic indicators are concerned – are scarce and while various indices and rankings on top domains across the world do exist, they are generally difficult to make sense of from a perspective of representativity.

No doubt, mapping the heterogeneities of online platforms in different languages and countries across the world is a crucial issue that needs to be more thoroughly addressed by the research community. This is also why this report has been compiled. Although social scientific researchers worldwide are increasingly realizing the power of online data, its terrain is complicated by ethical, legal, financial, and technical issues related to access and availability, representativity and replicability as well as ownership and privacy of online text content. This is an inevitable reality for many researchers, and while we are currently forced to accept certain limitations to our work, it should not discourage us from seeking new frontiers to study.

1.3 Pragmatic considerations: finding reasonable heuristics when dealing with web-mediated text

The general intent of this report is **to guide researchers who are interested in making country-comparative studies** and doing so by using internet-mediated text – a resource that is often provided by commercial data providers. From

a political-science perspective this is highly promising, since material from internet-based resources – both social (user-generated) and editorial (popular news media) – would, hypothetically, be a useful counterpoint to text data from, e.g., WWS and ISSP surveys. It would be nice to be able to use internet-mediated data for as many countries as possible – and, ideally, one would like to be able to assess the quality of that data. But seen as we all know how noisy and fragmented online-mediated text is, what pitfalls are there? What particular types of data would less suitable, and what data would be more suitable for comparative approaches? A humble starting point would be to provide at least some minimal detail, answering to basic descriptive questions, e.g.: Which Web-based editorial sources appear to be popular in each country, and which social media appear to be used in each country? Moreover, since access to such data is generally premised on the current ecosystem of application programming interfaces (APIs), allowing for certain institutional access, primarily on a strictly commercial basis, the supply side of such data is heavily skewed towards commercial access, prompting questions such as: Which data providers cover which platforms, and which languages? How much available data are there for each language, and what is the data quality like?

Even at first glance, some observable tendencies can be gleaned from our general overview. Interestingly, the relative popularity of various news sources can be compared, and in countries with a lot of political polarisation and debate, news sources with pro-government bias can be compared with those who are opposed to government or claim a neutral stance like in, e.g., the Hungarian news ecosystem. One can also glean **what relative impact Web-based online news media seem to have, relative to overall populations** – which reveals how weak and fragmented such news sources are, relative to other media in, e.g., sub-Saharan African countries, while, in the Nigerian context, at the same time revealing the relative abundance of microscopic news aggregation websites.

In addition to this section on Web-based news sources just mentioned, we include a section on **data providers**. Here, a descriptive account of the contemporary market ecosystem of Web scraping, aggregation, bundling, and analysis is presented. We know that the different corpus sizes for different languages would differ vastly; there is a lot more Web-mediated text available in English or Spanish than in Urdu or Tamil. To be able to assess such differences in a more cohesive way, a heuristic for rapid quantitative assessment is provided. While a lot of the providers in question do not provide full disclosure of their sources or give full access to their databases, this opacity does not restrict avid researchers from using makeshift measures to assess their relative corpus sizes. By, e.g., counting the prevalence of common stop words, one can assess corpus sizes and compare between different languages from the same provider. If one quantifies this relative difference in corpus size per language that different commercial providers deliver, one can see that some languages are overrepresented in comparison to their

actual popularity as spoken languages in the world (L1 and L2). This is of course attributable to the available supply of Web-mediated publications in each language, to begin with. Also, the geographic location of the provider seems to matter: Israeli-based providers give an overrepresentation of Hebrew content in their offering, while Swedish-based providers give an overrepresentation of Swedish content.

Another significant differentiation in bulk data like this, has to do with the **validity of the sources** found in each category. Considering the commonly occurring labels of "social" versus "editorial" text sources, arguably the most problematic category is the former one – "social" (i.e., user-generated) text. This is to do with both a matter of *definition* (what ought to count as user-generated text in the first place?) and *access* (on the open Web, there are fewer popular "social" web forums left, and those who remain are sparsely populated, since so many users have migrated to, e.g., Facebook-owned platforms, to which data providers and researchers are generally denied access). However, the "news" category is also problematic, in that providers routinely seem to cover a lot of heterogeneous sources – and while their initial filtering is generally adequate, often capturing many of the large news URLs in each language, it is far from perfect, and some inadequate sources might also be included in each corpus.
Hence, some kinds of quality assessment and filtering are recommendable, for researchers who want to attain validity in their source data. What was noted, as we observed the range of URLs on offer in the typical corpuses provided, was that the prevalence in terms of items (frequency of unique documents from each top-domain URL) followed a Zipfian (power-law) distribution, with a few very frequently occurring URLs and a long tail of much more obscure URLs, each tallying a very low frequency.

Using Swedish-language corpus text, from the same corpus as in the Lexicon (Dahlberg et al. 2021b), we tried two approaches – one where we simply split the corpus, at a specific threshold of item prevalence (2,000 documents, to be exact) and one where we, after having manually gone through all URLs that had more than 100 occurrences in the corpus, manually selected which URLs to include. Through using machine learning, we got results that indicated that the language in the documents that were weeded out qualitatively differs from the language in those that were included. This was not surprising. It confirms what would also be apparent through manual selection and analysis: The sites deemed invalid seem to contain language that is structurally quite different from the valid sites. This indicates that one should not include exactly all the data that is made available through commercial providers, but rather make a first rough selection, arguably "cutting off the long tail" at some appropriate point, when working with data like this. By not including the more obscure content in the long tail, one will be surer that one's data is valid and true to its categorisation as editorial news content.

The above, pragmatic considerations hint at something rather fundamental to any researcher of noisy, heterogeneous, internet-mediated data. As many scholars and non-academic pundits have noted, writing about the so-called 'big data' paradigm,[1] such data is known for its "three Vs": *velocity*, *variety*, and *volume*. In a stringent, academic overview of the actual types of data to be found bearing this label of 'big data', Kitchin & McArdle (2016) have shown that there are indeed multiple forms of such data, and not all of it needs to fulfil all the definitional factors enumerated. It is indeed possible to use small datasets, some of which neither being large in volume nor in variety, but whose ontological features are still to be considered as 'big data' in type.[2] That being said, if we are to contrast our "found" internet-mediated data at hand with, for example, survey data, it is reasonable to point out how the dialectical opposite of 'big data' – namely, 'small data' – must by definition be data that has been "produced in tightly controlled ways using sampling techniques that limit their scope, temporality and size, and are quite inflexible in their administration and generation" (Kitchin & McArdle 2016: 2).

> While some of these small datasets are very large in size, they do not possess the other characteristics of Big Data. For example, national censuses are typically generated once every 10 years, asking just c.30 structured questions, and once they are in the process of being administered it is impossible to tweak or add/remove questions. In contrast, Big Data are generated continuously and are more flexible and scalable in their production. For example, in 2014, Facebook was processing 10 billion messages, 4.5 billion 'Like' actions, and 350 million photo uploads per day, and they were constantly refining and tweaking their underlying algorithms and terms and conditions, changing what and how data were generated. (Kitchin & McArdle 2016: 2)

The same article (Kitchin & McArdle 2016) provides a very useful chart of the distinguishing features of conventional survey data, compared with administrative data and 'big data'. For our purposes in this report, the distinguishing features that we will observe are the following:

- Survey data tends to be purposefully designated and structured for pre-defined statistical purposes, ex ante – whereas found internet data ('big data') tends to have emerged organically (user-generated) or through

[1] While the etymology of 'big data' can been traced to the mid-1990s (Kitchin & McArdle 2016), the term was popularized in public and scholarly discourse in the years 2012–2014. See e.g. https://trends.google.com/trends/explore?date=2011-07-17%202021-08-17&q=%22big%20data%22,%22machine%20learning%22,%22artificial%20intelligence%22

[2] Take, for example, small datasets of tweets extracted from Twitter, where the sampling might have been very controlled and strategic, but where the ontological nature of the tweets in question still bear the hallmarks of big data: they are relational, spontaneously generated, polyvalent, polymorphous, and playful; they will doubtlessly contain a considerable amount of information that can only be understood in context.

planned administration *for different purposes than statistically representative sampling* (often highly commercial purposes). The structuring of found internet data often mirrors this fact, as it is often structured to be used by advertisers, for example.

- For survey data, conventional statistical methods (for assessing, e.g., representativity) are available, whereas they are not always suitable for found internet data. For the former, representativeness and coverage are almost always known by design, whereas, for the latter, representativeness and coverage are difficult to assess – but here, this report will endeavour to present a few useful practical heuristics that might help to at least roughly assess this, to a better extent than what the data might present at first glance. One such heuristic is to use Web traffic data for source URLs as a proxy for popularity.

- While survey data is almost always predefined and national in scope and ambition, as regards its sampling and target populations, found internet data is generated through platform users populating a platform, or users utilising a service, pass a sensor, contribute to a project, etc. Many of these emergent populations might be transnational, which – interestingly – means that language used becomes a more obvious sign of quantitative coherence than face nationality.

- Survey data generation is almost always slow and costly, and its known biases often low. Found internet data is potentially much faster and less expensive to come across; alas, its potential biases are often unknown or likely to be biased in various ways (some of these biases potentially disastrous). An important feature that will be noted in this report is that the current refusal of major platforms like Facebook, Snapchat and TikTok to be adequately transparent and helpful – these companies do not provide academic researchers data access at all – means that significant biases might be introduced, as the data that does get to be accessible to researchers might no longer represent the significant majorities of internet users.

- Survey data would generally have specified, clearly defined terms of use that are amenable to academic research, while the opposite could be said about found internet data; most major platforms severely restrict access to data (for reasons of user privacy and business strategy) and the data that is commercially accessible tends to be very recent, as older data has much lower use value to data vendors. Longitudinal 'big data' might therefore be hard to come across. The accessibility of found internet data is premised either on significant technical competence for manual collection or scraping (common, e.g., in data journalism) or, otherwise, on individual contracts and terms of service – and the terms of use can be subject to sudden

changes (vendors might change owners, become insolvent, have radical changes in their supply chains, or simply change their terms of service for undeclared reasons) or might lack clarity regarding specific uses that might be of interest to academics but not to the vendors' general customers.

For many of these reasons, we urge those readers who self-reflexively realise, at this stage, that they hold on to the commonly occurring modernist desire for perfect and complete datasets, to leave such desires at the door. In the world of user-generated, internet-mediated, found data – especially if one gets such data from vendors who have repurposed such data for commercial gain – one must start thinking much more pragmatically. We ought not to believe that this data will ever be comprehensive, or that it is akin to a rendering of the full extent of thoughts and opinions of the many – such hopes are idealistic figments of the modernist mind, desiring some final and total mirror of reality. Rather, we should observe this data as fraught with several troubling shortcomings and biases – but, nevertheless, indicative, in the main, of some broad tendencies and temperaments found in the leviathan of public opinion. The challenge is to understand and critically assess *which* tendencies and temperaments would be likely to register in such data, and reliably and validly so. This report will indicate that at least a few features could realistically be extracted from such internet data:

- While found internet data might not be ideal in terms of representativity of what issues, topics, and opinions are salient within the target population, they are nevertheless likely to be representative of the structural linguistic properties of each language 'in the wild' (with the proviso that what is sought are the relevant strata, or expressive styles/genres of language, see below). While the issues, topics, and opinions represented would be highly specific to each forum, it is likely that language used in online user forums would, to a significant degree, be semantically and grammatically structured like written language in other corpora.

- When dealing with internet-mediated text in bulk, the researcher ought to find a means for visualising or, by some other means, displaying the relative 'popularity weight' for each source; it might otherwise be misleading if a dataset of verbose text from some utterly obscure Web forum (i.e., large in data size due to document lengths) is taken to be more representative of a language than a dataset of more sparse text from an extremely popular forum (see Leech 2007 for an insightful discussion of this problem, from a linguistic point-of-view). To help researchers assess the relative popularity of sources, we suggest a method of identifying the relative volumes of Web traffic observed for certain websites (URLs) to more accurately assess the representativity of these sources.

- Domain knowledge is of course fundamental, in order to assess validity and legitimacy of the source in question.

- The output presented by supply-side vendors of internet-mediated text data is variegated, in that some languages are significantly more frequent in volume (and, expectedly, also variety and velocity) than others. This is important to note, when comparing datasets for different languages: smaller volume datasets increase the need for quality control of the data contained in each dataset, since local sampling errors and biases pose a greater risk of affecting the overall corpus for small datasets, compared to more robust, larger datasets.

Of course, demands for rigour and validity should always be maintained and maximised, whenever possible, but one needs to understand how demands for computability that are inherent to the highly inductive and deductive statistical methods offered by technology, engineering, and mathematics-oriented sciences (STEM) often clash with the sprawling and complex nature of society. As this report will show, this problem of incompleteness, noise, and biased data is further complicated by the severely restricted access afforded by dominant market actors like the Facebook corporation, and the commercial nature of those data that do get to be accessible. To any quantitatively inclined researcher, this is frustrating: Science could potentially be able to detect a whole host of interesting things that STEM-based methods enable researchers to find, if practical conditions were met. But the STEM-based methods can only happen in either an open academic environment, where rigour can be maintained but, alas, access is severely restricted – or in closed, corporate-controlled environments, where access is made possible, but researchers are bound by nondisclosure agreements to censor themselves and not share findings.

Further reading

It is recommended to read this report alongside a short reading list of academic texts that address similar issues:

- Allen, J., Mobius, M., Rothschild D. M., & Watts, D. J. (2021). Research note: Examining potential bias in large-scale censored data. *Harvard Kennedy School (HKS) Misinformation Review*. DOI: 10.37016/mr-2020-74
- Boullier, D. (2017). Big Data Challenges for the Social Sciences and Market Research: From Society and Opinion to Replications. In: F. Cochoy, J. Hagberg, M. Petersson McIntyre, and N. Sörum (Eds.) *Digitalizing Consumption: Tracing How Devices Shape Consumer Culture*. 20–40. Trans. Jim O'Hagan. London, New York, NY: Routledge.
- De Bolla, P., et al. (2019). Distributional Concept Analysis. A Computational Model for History of Concepts. *Contributions to the History of Concepts*, 14(1):66–92.
- Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*. DOI: 10.1177/2053951716631130
- Lewis, S.C., R. Zamith, and A. Hermida (2013). Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods. *Journal of Broadcasting & Electronic Media*, 57(1): 34–52. DOI:10.1080/08838151.2012.761702
- Mellon, J. & Prosser. C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, 4(3). DOI: 10.1177/2053168017720008
- Ruths, D. and J. Pfeffer (2014). Social Media for Large Studies of Behavior. *Science*, 346(6213): 1063–1064.

2. Challenges of validity and representativity

2.1 General challenges with digitally mediated text from a social science perspective

In this chapter, we will go through some of the more general caveats of dealing with **language data that are said or thought to be "representative" of general publics**. In social sciences, this is a huge topic, potentially unmanageably so. It's therefore with the utmost respect for the practitioners and theorists of numerous fields within social sciences that we will proceed, and only scrape the surface of this potentially inexhaustible topic.

Let us begin with one key observation from journalism and media studies, namely that editorial prose has its own dynamics and conditions for production, meaning that particular logics – e.g., in particular, **"news media logic"** (cf. Mazzoleni 2008) – will condition what is written and, hence, represented. It is generally assumed that "specific news content characteristics (such as conflict, personalization, negativity or scandal) are increasingly deployed by journalists to secure publicity, something that underlines the market and audience orientation in mass media communication" (Karidi 2018: 1237). Similar logics must be assumed to pertain also to newer media formats and genres, such as mass-appeal content designated to be "virally" shared across social media platforms (Chadwick 2013, Thomas & Cushion 2019), with the important caveat that such "social media logic" (van Dijck & Poell 2013) entails some additional factors, such as programmability, (metric) popularity, connectivity, and datafication.

It is, as such, clear that certain industrial and technological logics help shaping online editorial media sources, so that certain frames, genres, and conventions are seen in the *prima facie* text generated. It is our duty to call attention, early on in our analysis, to the historically contingent nature of discourse as episteme (that which is written and stated as "true" or "valid" about social reality at any particular point in time), and that future historians, when looking at the prose produced in our time, will surely be able to identify numerous generalities that are not even registered by us living at this time. In other words, any researcher ought to humbly accept that a totally comprehensive, exhaustive account can never be attained. It is, by the same logic, impossible also to fully account for what the alleged "real distribution" of opinion among the living public would actually be. The best we can do, from the perspective of quantitative analysis, is to map the observed distributions somewhat accurately in one domain/sphere next to those registered in another domain/sphere and see what the potential commonalities and differences are. This is, in a nutshell, what the LES project seeks to do.

Not only do we have to contend with the classic theory of "news logic" to begin with, we also have to humbly admit that representativity can, in itself, be

thought of in two rather distinct ways: On the one hand, we have representativity as the **identarian categorisation of "who gets to speak"** while we, on the other hand, have the **statistically distributional view of representativity**: "Are the distributions found in one sphere corresponding to the distributions in another sphere?" In what follows, it is this latter mode of representativity that will be addressed, while the former mode is arguably more commonly understood through qualitative and more ethnographically and anthropologically oriented approaches.

Lastly, before we delve into the challenges at hand, we must note that also when it comes to distributional representativity, this will, in what follows, be broken down into two methodologically distinct sub-categories, namely the question of **discursive representativity** (what topics and tonality/sentiment?) versus **structural linguistic representativity** (intralinguistic properties of language).

With that being said, let us now turn to one of the key challenges that this report is designed to address: **How to manage large data collections into something socially and culturally revealing? How to convert large amounts of information into something equitable and useful?**

Numerous researchers, within fields such as media and communications, journalism studies, computer-aided linguistics, and political science more broadly have grappled with these issues. In media and communications studies, Moe & Larsson (2012: 118) and Rogers (2019) have, for example, noted some of the more principal challenges that digital methods tend to throw up, from a social-science perspective. While digital source material can be seen as indicative of nothing more than its own **face appearance** (i.e., answering to an idiographic approach), researchers are often interested in *something more* than the mere appearances of online aesthetics, the characteristics of specific media or of the visible patterns of usage. Moreover, texts are also almost always used as technologies of **social normalization** (Weiss & Wodak 2002). Normative positions can move, collectively in (nationally and/or linguistically bounded) societies, in terms of what appears to be publicly acceptable at one point in time, to what is deemed conventional and/or allowed to be said at a later point in time (Wodak 2018).

Now, say if we wanted to use text data compiled from online sources as specimens of – for example – larger, more popular discourses in a particular society, how to approach this data? As was noted above, content that is found online should always be read with an eye to its specific conditions for emergence. One should not rush too quickly to conclusions out of mere face validity, or over-interpret the significance of texts – but, at the same time, one's found specimens ought to be seen as indicative of *something* (i.e., answering to more nomothetic ambitions).

2.2 Challenges from the perspectives of quantitative media content analysis, and, respectively, corpus linguistics

2.2.1 Challenges from the perspective of quantitative media content analysis

How has opinion research (including journalism and media studies) approached the challenge of validity of online-mediated text?

Within opinion research, it is long-established that editorial prose, while indicative of the goings-on and public opinions in a society, cannot be taken "at face value" as quantitatively representative of public opinion, as if there was a 1:1 correspondence between the quantity of arguments in the press, and the actual prevalence of the same argument within the population. Nevertheless, quantitative analysis of media content (e.g., Krippendorff 2013) is a central component of opinion research, since it would otherwise be hard to **quantify popular opinion and normative discourses in societies**. Another useful definition of quantitative content analysis is offered by Riffe, Lacy, and Fico (2014: 19): "the systematic and replicable examination of symbols of communication, which have been assigned numeric values according to valid measurement rules, and the analysis of relationships involving those values using statistical methods, to describe the communication, draw inferences about its meaning, or infer from the communication to its context, both of production and consumption."

What is important to note, before we move along to a discussion of online text resources, is that while media content analysis has always occupied itself with the question of **sample representativity**, the issue whether that entire body of text (the corpus) that the sample is thought to represent is *in itself representative of the larger population* is a much thornier, in many ways contested question, as Klaus Krippendorff points out in his primer on content analysis:

> A stereotypical aim of mass-media content analysis is to describe how a controversial issue is "depicted" in a chosen genre. Efforts to describe how something is "covered" by, "portrayed" in, or "represented" in the media invoke a picture theory of content. This approach to content analysis decontextualizes the analyzed text and thus […] conceals the researchers' interest in the analysis, hides their inferences behind the naive belief that they are able to describe meanings objectively while rendering the results immune to invalidating evidence. Consider common findings of political biases, racial prejudices, and the silencing of minorities on television as such issues. Although counts of evident incidences of such phenomena can give the impression of objectivity, they make sense only in the context of accepting certain social norms, such as the value of giving equal voice to both sides of a controversy, neutrality of reporting, or affirmative representations. Implying such norms hides the context that analysts need to

> specify. Unless analysts spell out whose norms are applied, whose attitudes are being inferred, who is exposed to which mass media, and, most important, where the supposed phenomena could be observed, their findings cannot be validated. Berelson and Lazarsfeld […] noted long ago that there is no point in counting unless the frequencies lead to inferences about the conditions surrounding what is counted. For example, counting the numbers of mentions of Microsoft or AIDS or the term road rage over time in, say, the New York Times would be totally meaningless if the observed frequencies could not be related to something else, such as political, cultural, or economic trends. That something else is the context that lends significance to quantitative findings.
> (Krippendorff, 2013: 34)

As was already noted by reference to the theory of "news media logic," no-one who conducts a study of editorial text in a national leading newspaper would be under the illusion that, while one's study is careful to select a representative sample of that specific corpus, these accumulated editorial articles (in their totality) would be automatically representative of a particular population, or even a particular ideological contingent within that population.[3]

With the advent of large-scale, rapidly growing collections of digitally produced online prose, this notion has taken a rather fascinating turn. The advent of the internet seems to be joined by an imaginary notion, that presupposes that text on the internet must be somewhat truer to actual civic reasoning, as an imagined outpouring of privately held beliefs and opinions among the larger population, as the majority of citizens use the internet, and large groups of people congregate in online forums where they can, with little effort, verbalize opinions and sentiments, in formats that are publicly findable and traceable.

However, it appears to be very idealistic to presuppose that such outpourings would be unproblematically indicative or indeed representative of the actual beliefs and opinions held by the population at large.

> Social media analyses may [...] be more ecologically valid than traditional approaches. Unlike approaches where responses from participants are elicited in artificial social contexts (e.g., Internet surveys, laboratory-based interviews), social media data emerges from real-world social environments encompassing a large and diverse range of people, without any prompting from researchers. Thus, in comparison with traditional methodologies, participant behavior is relatively unconstrained if not entirely unconstrained, by the behaviors of researchers.
> (Andreotta et al. 2019)

---

[3] Historically, newspapers as we know them today emerged out of political pamphleteering (17th C) and legacy papers tend to have some kind of (more or less openly stated) ideological affiliation or leaning in their editorial sections, as media historians (Briggs & Burke 2020, Curran & Seaton 2018) have shown.

Nevertheless, social media data has several attributes that make it hard to reliably monitor, capture, measure and assess: It is interactive, ephemeral, dynamic, and – at least on an aggregated level – massive in volume (Andreotta et al. 2019). Social scientists have noted that the material features of social media give rise to a range of social media logics: programmability, popularity bias, connectivity, and datafication (van Dijck & Poell 2013).

In the introduction, we outlined several features of the contemporary, digitally generated, ever-larger datasets. As such data becomes increasingly prevalent, media scholars are trying to find different ways of using computational methods as part of their overall analysis. One of the most cogent summaries of the issues at hand is Zamith & Lewis (2015), who outline how the traditional (manual) approach to conducting a content analysis is being reconfigured and try to assess what is gained and/or lost when turning to computational solutions. One of the key conclusions that they make is that computational methods are particularly valuable when variables are readily identifiable in texts and when source material is easily parsed, but that manual methods still are appropriate for complex variables and when source material is not well digitized. They argue for a hybrid approach, where modes are "effectively combined throughout the process of content analysis to facilitate expansive and powerful analyses that are reliable and meaningful" (p. 307).

Of course, there are numerous sources of bias that can affect data fetched from the internet. Ruths & Pfeffer (2014) provide a good checklist for large-scale academic studies of human behavior in social media. Several issues prevalent in social science studies using large datasets are highlighted and addressed with supplementary strategies. With the advent of methods based on the collection and analysis social media data, aided by machine learning and network analysis, the authors argue that researchers must consider several pitfalls that have emanated with these developments in the field. In order to refine the quantitative analysis and representation of populations and their behaviors, researchers should address **population bias** in sampling, **dependence on proprietary algorithms** which are subject to change and irreproducibility, as well as the **limited access to platform-specific data** that they usually entail. Other issues with data from social media platforms are that such data often contain information from profiles that are **hard to assign** to specific populations of interest – many internet users do not label themselves, and therefore they cannot be quantified. There is also a massive number of **spam content and fake profiles** on most platforms, as well as promotional accounts that are hard to filter out from results. **Certain aspects of human behavior are also not translated** into the data that is collected by platform corporations; neither the activity that precedes actual searches nor full chains of retweets are stored, which could distort and omit certain aspects of online social behavior. The **terms and conditions** on these platforms have also been understood to have a serious impact on the nature

of academic research – sharing of datasets is often prohibited by the actors, as we will see below (2.4.4), preventing comparative analysis and evaluation of computational methods.

In their attempt to answer these issues, Ruths and Pfeffer encourage the sharing of methods at publication time, while also observing that the academic culture at large would benefit from the publication of negative findings and failed studies, in order to assess the extent that successful studies in this field rely on random chance. This could help map existing best practices, so that analytical bias can be managed better in further statistical research. Finally, the authors present a checklist of approaches, which include the quantification of platform-specific biases, biases of available data, and proxy population biases/mismatches. For methods, they prescribe the application of filters for non-human accounts, ask researchers to account for platform-specific algorithms, and to compare the results from other existing methods on the same data.

Similarly, Boullier (2017) has provided a schematic overview of what he sees as three generations of quantitative social science: census records of early modernity; sample-based methods of high modernity (relying on normal distributions of variations in populations); and the high-velocity, large-volume real-time indications of contemporary "big data" (no longer relying on normal distributions as such distributions are often Zipfian). In his schema, speaking of traditional statistical representativity is to rely on the second-generation, sample-based approaches of high modernity. When applying such approaches to contemporary "big data," one is combining historically distinct approaches; arguably, assigning information from profiles to specific populations of interest (as per Ruths & Pfeffer above) is to also include the first-generation approach, mapping data onto a census as if it were.

What is more, in the overview that is conducted below, a striking tendency is observed, that the two categories that commercial providers of online-mediated text are offering is suffering from a skew, in that while "editorial" text is thriving and remains largely accessible through these providers, so-called "social" text (harvested from civic online interactions) seems to be getting gradually rarer to come by, as fewer and fewer online forums remain active in the wake of a handful of gigantic social-networking platforms sweeping the world. Out of these large, proprietary platforms, the largest one – Facebook (owner of WhatsApp and Instagram) – has shifted to a model where the platform giant unilaterally stipulates the terms-of-trade, offering zero to little raw content to be made available. Compare the Web forums of the past that were publicly accessible and overseeable with today's situation where gigantic corporations run walled gardens, in which enormous volumes of public discourse are hosted, but access and transparency is completely lacking, leading scholars to label these infrastructures as "black

boxes" (Bucher 2018, Pasquale 2015) Twitter runs its own data sharing interfaces, offering bulk data as a commercial service, as a complement to the data available through the independent data providers. What remains, are mainly blog posts and web forum text, and it is well known that such media services have become drastically less popular in many countries, as internet users flock to the large, transnational platforms instead. These tendencies are described in more detail below (sections 2.4.3 and 2.4.4).

This general sea change gives even more credence to the position advocated in this guide, namely that while online text remains a highly useful resource for scholars of public opinion, across the globe, in all sorts of linguistic and political contexts, researchers must hedge the impulse to see such text as automatically "representative," by adhering to several caveats.

What is therefore necessary is contextual understanding of the forums and sites in question, much like the understanding (today often taken for granted) that random samples of newspaper prose should not be seen as representative for much more than the corpus itself. Once quantitative validity and reliability is established, in terms of the sample's relation to the corpus, much more epistemologically challenging questions must be asked:

> The foregoing suggests that purely descriptive intents, manifest in claims to have analyzed "the content of a newspaper," to have quantified "the media coverage of an event," or to have "found how an ethnic group is depicted," fail to make explicit the very contexts within which researchers choose to analyze their texts. Content analysts have to know the conditions under which they obtain their texts, but, more important, they also have to be explicit about whose readings they are speaking about, which processes or norms they are applying to come to their conclusions, and what the world looks like in which their analyses, their own readings, and their readings of others' readings make sense to other content analysts. Explicitly identifying the contexts for their analytical efforts is also a way of inviting other analysts to bring validating evidence to bear on the inferences published and thus advance content analysis as a research technique. (Krippendorff, 2013: 34–35)

As regards the terms emphasized by Krippendorff, it is instructive that researchers have **contextual knowledge** of the societal settings in which their research is conducted, understanding of the **motivational concerns of social agents involved**, and **situational knowledge** about their research objects. Evidently, one such source of understanding comes from local knowledge of the countries or languages involved – i.e., the relative importance and significance of, e.g., specific forums or discourses in that setting. Moreover, it is instructive to have local knowledge about the vendors in question – their purposes and surrounding institutional landscapes (i.e., competitors, business incentives, etc.).

It is therefore a guiding principle for this study that different ways of obtaining context must be sought, in order to place corpora and samples in their right circumstances. By mapping the landscape of commercial providers (vendors) and the distribution of sources in different countries and languages, we hope to be able to compile a set of observations that should aid researchers in such attempts.

2.2.2 Challenges from the perspective of corpus linguistics

How has corpus linguistics approached the challenge of representativity, and how can corpora be designed to be representative?

Representativeness is a concept that is foundational to sociological research. In all descriptive statistics that are intended to characterize the overall population, it is crucial that sampling is stringent and designed so that representativity is maximized. In sociolinguistics, terms like "corpora" and "populations" are interchangeable with equivalent terms in opinion research. Moreover, the term "strata" is introduced, as we shall see below.

It is not realistic to expect any corpus to be representative for the entirety of a human language in all its written and spoken instantiations. Within larger corpora, **researchers therefore elicit specific *strata*** (Biber 1993), so as to be able to weigh, focus, or highlight particular types of texts – e.g., samples of selected genres, or types of publications within the population. In essence, such strata constitute groupings and categorizations that are intended to characterize populations by, crucially, being useful samples. It is important, of course, that the selected strata are proportional to their actual occurrences among the population: "If a stratum which makes up a small proportion of the population is sampled heavily, then it will contribute an unrepresentative weight to summary descriptive statistics" (Biber 1993: 247).

Strata, consequentially, refer to subgroups within the target population (e.g., genres) which can each be sampled using random techniques. Biber (1993) exemplifies by noting the attempts, by designers of large corpora, to make exhaustive listings of the major text categories of published English prose, and then include all of these categories in the corpus design. "Random sampling techniques were required only to obtain a representative selection of texts from within each subgenre. The alternative, a random selection from the universe of *all* published texts, would depend on a large sample and the probabilities associated with random selection to assure representation of the range of variation at all levels (across genres, subgenres, and texts within subgenres), a more difficult task" (Biber 1993: 244, emphasis added).

In language corpora that are intended to represent whole languages, like the ones in Biber's examples, a slightly different notion of representativeness is required, where proportional sampling is less appropriate. In conventional linguistics, researchers are generally interested in the full range of linguistic variation existing in a language, which means that language samples are required that can include such variation. Still, "a proportional language corpus would have to be demographically organized, because we have no *a priori* way to determine the relative proportions of different registers in a language" (Biber 1993: 247). At the same time, statistical regularities appear, regardless of sample sizes, such as "Zipf's law" (cf. Newman 2005) – which indicates that, as a general tendency, word occurrences are distributed according to mathematically predictable patterns.

In the LES project, this is however not the primary concern. We are interested in observing and trying to quantify the semantic differences in terms of *distributions of neighboring tokens specific to specific languages*, so that important terms for political science (e.g., "democracy") can be more easily compared across languages – with the important caveat that our observations are restricted to open-ended answers in social science surveys and open-ended text found on the Web. The strata in question should therefore be regarded as the one of **politically/societally oriented free text survey answers** versus, on the one hand, **politically/societally oriented editorial prose published on the internet**, and, on the other, **politically/socially oriented user-generated forum text on the internet**. We are under no illusions that those sources would be readily representative of larger populations in a 1:1 fashion. Rather, the project aims at improving the relative comparability across languages, *within* these already established registers. In other words, we are interested in finding the specific *relational characteristic* of terms like "democracy" in (a) survey answers, (b) editorial prose, and (c) citizen-generated online text in different languages – e.g., Spanish, as compared to Russian, Indonesian, or English.

We nevertheless intend our respective corpuses to be as representative of each country's or language community's demographic composition as possible. Among linguists, the objective of representativity is also what would typically distinguish a corpus from a mere archive. While an archive is an ordered, but to many extents arbitrary collection of texts, "a corpus is designed to represent a particular language or language variety whereas an archive is not" (McEnery et al. 2006: 13). Hence, a key consideration for corpus design is proportionality. Some degree of selection and filtering is always required. The question is how to make the sampling as representative as possible.

Sampling frame refers to the overall choice of what archival sources to peruse, both in diachronic (chronological) terms, and in terms of synchronic choices (securing the provenance of texts; cf. Leech 2007: 145). Always, when choosing and

evaluating a sampling frame, considerations of efficiency and cost effectiveness must be balanced against potential rewards as concerns the granularity and degree of representativeness (Biber 1993: 244). Here, it becomes obvious that corpus representativeness itself is always both a theoretical (Halliday 2005) and methodological construct (Leech 2007), where certain design choices are always made. The researchers actively choose to include certain sources and exclude other ones. Interestingly, this is also the case when scraping is automated; manual choices are always required anyway – regarding, e.g., which top-domain URLs should be included or not, before the computer begins scraping everything from that URL. Raineri & Debras (2019) note that "multiple theoretical, methodological and practical questions are raised by the issue of corpus representativeness." What to choose as groupings and categorizations (strata) in one's corpus is an act of interpretation on the part of the corpus builder, and it could be argued that genres are never naturally inherent within a language, always attributed and pointed out as such by someone. Genre groupings are, in effect, produced by the theoretical perspective of the linguist who is carrying out the stratification. While questions like these are by no means new, they continue to draw attention in current research (e.g., Gray et al 2017).

Optimizing for reach (popularity) or frequency (verbosity of prose)?

When we relate this to the online milieu, the logical conclusion is that the corpus design should strive to mirror the actual textual universe that it is a sample of, so that corpus items are somewhat proportional to either the numbers of **instances of reception** (i.e., manifest popularity of mass media in terms of reach/circulation) or the numbers of **instances of production** (i.e. manifest popularity in terms of numbers of aggregated individual postings) in the larger population. Arguably, the latter is a problematic metric, however, since it is nowadays obvious that forums and message boards are often flooded with comments that are only representative of certain individuals with strong opinions and specific agendas, never the entire universe of opinion outside of the forum as a textual space. Therefore, the first-mentioned metric – manifest popularity of certain sources, in terms of instances of reception (i.e., how many exposures the sources would have generated) – would be a slightly better, but by no means perfect, proxy for popularity. However, the actual popularity of a site or publisher can rarely be gleaned from the data available.

Consequently, many of our research design choices come down to an epistemological discussion of *what should count as representative*, in terms of language specimens. Should weight be assigned to popularity, so that a short paragraph that has been read by one million people would be thought to count as more relevant than a paragraph that has only been read by one person? As Leech (2007) once noted, such approaches have more importance than just academic point-scoring.

On the internet, as we shall see, the quantitative differences are often as stark as this, and it really poses a problem – not least for language models used in machine learning.

On the other hand, we find that among many historically inclined linguists, and among those sociologists and historians of literature who care for corpus-based methods, regular expressions are often gleaned from corpus text that might be comprised of literature and prose that might never have seen large readerships, in terms of audience reach and large publics. Biber (1993: 248) has noted that, in corpora where each text is represented as one unit, "registers such as books, newspapers, and news broadcasts are much more influential than their relative frequencies indicate." The variable length of texts plays a role in determining the likelihood of a text being representative of the distribution of discourse in populations: "Thus a tabloid newspaper such as *The Sun* (in the UK) contains fewer words per issue than a broadsheet newspaper such as The Independent. This should give *The Independent* greater sampling privilege which would partially offset the smaller circulation of that paper" (Leech 2007: 140). The language specimens that are fed into machine learning algorithms often come from compilations of prose, where the singular articles might have been published in small-circulation journals, or in books with very small print runs. Likewise, a lot of online text specimens are harvested from Wikipedia, where, equally, a long-tail distribution of popularity reigns supreme (as everywhere on the internet) and where, consequentially, most text specimens are from pages that have had very few actual page visits.

But does "few page visits" equal "less representative"? Not necessarily when it comes to the **internal structure of language**;[4] the highly rule-based nature of written language and the prevalence of clear writing norms across communities mean that the variations in purely linguistic style between a large-circulation text and a small-circulation text might be insignificant. This is not to say that there do exist such stylistic variations (e.g., sociolects, slang, vernacular, etc.) but it is not necessarily always the case that such potential variations should be concomitant with "audience reach" as a discrete parameter.

In the literature on linguistic corpus design, these two major types of representativeness are generally referred to as *target domain* and *linguistic representativeness* (McEnery et al. 2006). Target domain representativeness can also be called *external* (e.g., McEnery et al. 2006) or *situational* (e.g., Biber 1993) representativeness, while linguistic representativeness has been referred to as *internal* representativeness (e.g., McEnery et al. 2006).

---

[4] Leech revisits a famous distinction in linguistics, established by Chomsky (1987), between "E-language" (externalized language) and "I-language" (internalized language). This distinction complicates things even more, as it draws upon questions of cognition as well.

Gray et al. (2017) define target domain representativeness as the extent to which a corpus contains the full range of text type variability that exists in the target domain. "Target domain representativeness determines the generalizability of a corpus sample to a larger population of interest" (p. 1).

Linguistic representativeness charts the extent to which a corpus contains the full range of linguistic distributions that exist in the target domain. Linguistic representativeness determines the suitability of a corpus sample for answering specific research questions about specific linguistic features. Importantly, linguistic representativeness is inherently related to the linguistic level being investigated; the same corpus may, e.g., be representative of a common grammatical structure, but not of lexical distributions.

For corpora harvested from the internet, this constitutes a challenge. One needs to demonstrate the suitability of one's corpora, and one way to eliminate troubling discrepancies is to assess corpora by substituting different corpora for other ones. If we have data from one vendor, would another vendor, with somewhat different web sources in its corpus, yield similar or very different results? Does a corpus consisting of relatively obscure sources throw up different results than a corpus with manifestly popular sources? Ideally, this would have to be clarified in order to assess the impact that variability of sources would have.

For such purposes, renowned linguists like Mark Davies (n.d.) have recommended that "keeping texts intact and carefully documenting metadata regarding their source and characteristics allows the researcher to create a corpus sample that can be meaningfully stratified or described in many different ways." In other words, it is good to have access to parameters such as manifest popularity of different web sources, in order to assess the "orders of magnitude" of reach of particular websites. Indexing services, like e.g., SimilarWeb, Alexa, or ComScore can reveal whether web sources have millions of unique visitors or only fractions of a percent of that popularity.

Also identifying **the approximate numbers of speakers of a particular language**, versus **the manifest amounts of recorded text available in various corpora for each language** is a useful way to assess the adequacy of one's sampling frame for that particular language. Here, a test has been designed by us, where an approximate listing of relative sizes of languages in the world (counted in numbers of L1 and L2 speakers) is related mathematically to the estimated corpus sizes for each language, among the different vendors. This way, we can (a) get an assessment of the relative "rate of coverage" for each vendor and language, and (b) compare different vendors with each other, so as to deem which vendors would be suitable for which certain countries and languages (see below, section 3.2).

2.3 The overlaps between corpus linguistics and quantitative media content analysis

If we go back to the dilemma of **whether to see large-circulation web sources as more representative than small-circulation ones**, it is clear that this dilemma is not so much to do with stylometric variations of language (linguistic representativeness) as it is to do with external or situational representativeness (target domain), and the attendant differences of framing and discourse that can doubtlessly be noted in different online or offline milieus. Which synonyms tend to be used instead of other ones, in each specific stratum? Which specific argumentative claims tend to be linked to which other claims, in each stratum? It might be very likely that texts with large-scale circulation might contain higher degrees of elite reasoning, for example, while small-scale circulation texts might be "truer" to everyday vernacular of ordinary citizens. On the other hand, sources intended for large-scale circulation often tend to be quality-controlled by editors and fact-checked as, in most societies, such text tends to be subject to significant degrees of scrutiny (through more or less formal means), whereas text with small-scale circulation rarely has any stipulated requirements to adhere to norms of veracity and coherence, and might therefore be more prone to quack reasoning, or even outright conspiratorial and/or delusional nonsense – or, more commonly, machine-generated junk/spam intended to lure indexing sites and advertising algorithms.

Ultimately, text specimens that are known to have been published in a large-scale circulation context should be expected to have reached and resonated with more readers than texts from small-scale circulation contexts. This might not mean that any given specimen will be guaranteed to have resonated *well* with its audience – it might have been dismissed, misunderstood, or misread in any sense of the word but, over time, one should estimate that cumulatively, the specimens from a particular context would be *structurally attuned* to that context. It is not likely that a large, national newspaper would keep publishing texts that would continuously be dismissed or misunderstood by the vast majority of its readers; it stands to reasoning that a harmonious relationship would be sought between readers and writers over time – and that, as such, specimens from, e.g., large national newspapers would say something relevant about the discourses circulating among large groups of people in that society, at that point in time.

At the same time, we all know that the press system – indeed, all large-scale media with editorial and/or advertorial ambitions – has inherent biases, stemming from things like news values, audience maximization, and the appeasement of advertisers, owners, and/or shareholders (once again, "news media logic"; Mazzoleni 2008). This all leads to the fundamental questions that this section began with, questions that have been well-established in media and communications

scholarship and journalism studies for decades. These are core questions for opinion research since its very conception: Can editorial prose (news prose in particular) be said to be representative of its foundational society, given all the known, inherent biases to media representation?

This is too large a topic to cover in its entirety – but it is of utmost importance, since if we are in the business of compiling epistemological and methodological challenges with online text, we must respectfully address all the inherent contradictions and paradoxes to any inquiry that attempts to objectively say something about the representativity of written language. As it stands, we will pursue our task, but we shall note that several caveats will arise due to issues like the abovementioned.

2.4 Practical research challenges for linguistic explorations of societies through web-mediated text

2.4.1 Formats, genres, text types

It should be clear from this summary that conventional corpus linguistics are in many ways similar to quantitative content analysis in media studies, as special attention is devoted to distinctions between *genres*, *registers*, and *text types*, and the apparent conventions of such genres and types. Biber (1993: 244–245) uses the terms *genre* and *register* to refer to "situationally defined text categories (such as fiction, sports broadcasts, psychology articles)," and *text type* to refer to "linguistically defined text categories" (i.e., "shared linguistic co-occurrence patterns, so that the texts within each type are maximally similar in their linguistic characteristics, while the different types are maximally distinct from one another").

Translated to our present-day task of sampling online text, some key genres and categorizations should be listed:

Table 1. List of possible formats, text types, genres

Possible formats, genres, text types

- formats
  - headers;
  - body text;
  - user comments;
  - metatext (i.e., image captions);
  - repostings (i.e., aggregation of text originally published elsewhere);
  - metric indicators of popularity;
  - etc.
- text types (categories)
  - original reporting;
  - opinion;
  - user comments;
  - etc.
- genres (registers)
  - domestic;
  - foreign;
  - economy;
  - tech;
  - entertainment;
  - sports;
  - etc.

Here, we must maintain the proviso that there are significant challenges to do with boundaries sometimes being hard to draw and hybridity often occurring, and that for practical reasons, probably higher-level strata will have to make do for a lot of sources, i.e., scraping entire websites while knowing that what is captured are several of the abovementioned more granular categorizations. Once again, the goal is not to represent complete languages, only to make possible commensurability across countries and language communities, so that largely editorial media from several countries can be aggregated and compared, respectively to each other, in a way that is as appropriate as possible.

In defining the population for a corpus, register/genre distinctions take precedence over text type distinctions. This is because registers are based on criteria external to the corpus, while text types are based on internal criteria, i.e., registers are based on the different situations, purposes, and functions of text in a speech community, and these can be identified prior to the construction of a corpus. In contrast, identification of the salient text type distinctions in a language requires a representative corpus of texts for analysis; there is no *a priori* way to identify linguistically defined types (Biber 1993: 245).

Dialectical variety and informal vernacular expressions are unlikely to appear in written text.[5] Typically, when making content analyses of media text, the linguistic features of such text almost always tend to be of a "standard" variety (e.g.,

[5] In 1993, Biber stated that the state of the art of linguistics was that studies of dialect variation from a text-based perspective were very rare; dialect studies at the time "tended to concentrate on phonological variation, downplaying the importance of grammatical and discourse features" (p. 256, note 2).

normative, nationally ordained single dialects). Consequentially, the corpus design of editorial text would not show great variety in terms of dialect or vernacular; the differences would instead be on the level of conventions of collocated terms, a highly ideological and politically charged arena of discrepancy—take, for example, the political issue of how to refer to demographic categories of class, nationality, ethnicity, gender and so on. For example, does the text use formal terms like "asylum seekers" or "refugees," or does it use pejorative terms like "welfare tourists" or even "parasites"? Does the text use highly abstract terms like "collateral damage," gradually less abstract like "non-combatant casualties" or more concrete synonyms like "civilian victims"? It is on this level where stark differences are likely to occur between different genres, different types of media outlets, and different types of agenda-based reporting.

The above largely applies to editorial online text. What about the category of *socially* produced text on the internet, then? In broader linguistics, researchers are concerned with the variability of the multiple dimensions of speech (e.g., phonology and phonetics, prosody, gesture etc.); the range of variability in populations, in terms of situational (e.g., format, setting, author, addressee, purposes, topics) and distributional linguistic parameters (e.g., frequencies of word classes). Leech (2007) discusses this by reference to Chomsky's (1987) "I-language" (internalized grammar) and "E-language" (public languages, as manifested, e.g., in writing), and concludes:

> It is true that the Web gives access to a very wide range of genres, some of them well-established in the written medium, such as academic writing and fiction writing; others newly-evolving genres closer to speech, such as blogs. However, it is also true that the Web by definition gives little or no access to private discourse, such as everyday conversation, telephone dialogues, and the like. Searching with a search engine provides no access to spoken or manuscript data. There are major areas seriously underrepresented, if they are represented at all. It is also likely that certain varieties, such as academic writing, are overrepresented. The multi-media and HTML format of webpages is also likely to exercise its own constraints and preferences in the use of language. (Leech 2007: 144)

With such internet text, the problem is often one of chronology and of provenance: "It is obvious that the Web provides nothing like the exact comparability of text selection for different periods or different regions of the world. On the diachronic axis, it is even impossible to tell when a particular text or text extract was written; similarly, on the synchronic axis, knowledge of the provenance of a text is minimal" (Leech 2007: 145).

As we will see below, the commercial reality of data provision on the Web means that profit-motivated vendors who provide internet-mediated language data in bulk often have an incentive to ensure that this data is new and current, which

means that data older than 30 days is demoted, i.e., less commonly included in the services' offerings and, in effect, often harder to come by. Moreover, the timestamps, geolocation and language tagging of the data offered are generally repurposed from the original metadata of the HTML files, as they are published, and, while some manual curation is often made at the point of choosing to include particular sites or not *("Should news publication X be included in the data stream?"*), such curation is rarely made on a daily basis (*"Is each and every article from publication X double-checked for metadata and validity?"*), since the whole point of these vendor services is that enormous volumes of data can be made available on an automated basis.

2.4.2 Platforms, accessibility, and the evanescence of the Web

The so-called "platformization" of the web has been a key feature of the global internet since at least a decade. Scholars of media and communication (e.g., Tarleton Gillespie, Anne Helmond, José van Dijck, Taina Bucher, and other notable scholars) have observed how, increasingly, various internet companies have come to effectively build so-called "walled gardens" where user verification by means of usernames and login procedures are mandatory to at all participate and access online content. This is a trend that has been exacerbated by various factors, such as the need to generate profitability in the online realm (which explains the tendency among publishers to build so called "paywalls" around their content), and also the emergent app infrastructure that originated with the foundation of an app economy of mobile internet devices in the early 2010s. This tendency to "hide" online text behind a login has in effect made the online realm a sequestered, reterritorialized space, which drastically affects the availability of online discourse – for commercial providers, noncommercial researchers, and ordinary internet users alike.

In his typology of strata, Biber (1993: 246) notes that there are three types of setting that can be distinguished for printed matter: "institutional, other public, and private-personal." The problem, in 1993, was that no direct sampling frame could be said to exist for *unpublished* writing or speech. In 2007, when Leech wrote, the internet had been established as a popular arena for more than a decade, but the timing of his writing is interesting because at the time, the global internet stood at the cusp of the mass-scale platformization, changing the very nature of the internet considerably. Consequentially, thereafter we have seen considerable reterritorialization, corporate consolidation, and structural centralization (e.g., Jin 2013, Hindman 2018). Today, while there is a glut of digital, internet-based person-to-person and vernacular media, a significant challenge to researchers is that such communications, while legitimately private, are simply inaccessible except for inhouse researchers at the very corporations in question that run these

social networking services (Facebook, Google, Apple, Microsoft etc.). Central components that govern the field of technical and organisational interactions between platform architectures like those of Facebook, and its numerous external stakeholders are application programming interfaces (APIs), software development kits (SDKs), and reference documentation (Helmond et al. 2019: 124). Alongside the commercial contracts governing the legal aspect of platform ecosystem consolidation, these API and SDK functionalities have become gradually restrictive over time, so that neither commercial vendors nor independent researchers can freely access these services. In order to access such data directly from the platform corporations, normally one would have to sign away one's independence as a researcher.

Going through the metrics for popular websites (editorial, social) in various countries, it is obvious that in the early 2020s, the social web is basically a global oligopoly. Domestic social networking sites and forums did exist for a very brief time, but over the last decade they have almost all disappeared. Take Germany for example; a populous, rich country with a very large language community. Logically, it has dozens of thriving online editorial media. But if we are to regard domestic, German-origin social sites, they are almost all gone. The only relatively popular domestic site remaining in 2020 was Xing.com with ~8M monthly unique visits. Other, previously popular sites, like student-networking site studivz.net are practically defunct. Lokalisten.de closed in 2016; wer-kennt-wen.net in 2014. One of the providers in our overview of data providers (section 3.1.1 below) used to claim 30 million forum posts per day in their data stream a couple of years ago, while the same provider now claims 10 million forum posts per day. This is a significant change in volume, since one would otherwise assume that online content is expanding in frequency over time.

The last decade has seen a huge clearing of the global markets in social networking sites, primarily by US-based platform corporations, alongside a handful of Chinese (WeChat, Sina Weibo, Tencent, Douban) and Russian (VKontakte, OdnoKlassniki) platform corporations. For country after country surveyed, the leading social networking sites are Facebook and Facebook-owned services like Instagram and WhatsApp; Google-owned Youtube; Microsoft-owned Skype and LinkedIn, alongside American companies Pinterest, Twitter, and Snapchat.

Another issue was found, which was that some top domain URLs were categorized by the web scraping providers as being social forums, but at closer scrutiny many of these actually would more correctly qualify as marketplaces – like, e.g., German social networking site Xing.com, French business networking site Viadeo (available in English, French, German, Italian, Portuguese, Spanish, Russian), marketplace sites like Kijiji, Bobobo, Immobiliare, and Subito (Italy), Blocket (Sweden), or, for that matter, dating sites.

Another significant problem, when trying to access hyperlinks provided in the corpus material, is that the Web is inherently ephemeral. While documents might exist in the corpus, or through the provider's internal dashboard access to its database – they might be no longer available through the publisher's or host's public interface, or indeed to be found on the Web at all. Practically, this adds to the dilemma of not being able to subsequently cross-check archived specimens through the available public interfaces of publishers or forum hosts, long after they were originally harvested.

2.4.3 Platforms unilaterally restricting commercial access

By referring to a twofold narrative, **giant social media platform companies (Facebook in particular) are policing the means of data access for commercial data intelligence companies**: The primary reason stated being a care for user privacy, the secondary reason being business secrecy. As we have already noted, application programming interfaces (APIs) are important mechanisms to provide "a set of interfaces" that enables external web- sites, platforms, and apps "to communicate, interact, and interoperate with the platform" (Tiwana 2014: 6).

> Consequently, they allow third-party developers, such as marketing agencies, to build "on top of" a platform's core infrastructure, thereby extending its functionality. Relatedly, SDKs [software development kits] are important boundary resources that facilitate and streamline the app development process by providing developers with a set of software tools, developer libraries, APIs, documentation, code samples, and guides. (Helmond et al. 2019: 127)

In order to protect the privacy of its users, Facebook sets certain parameters for how third-party data providers can access these technical resources. Moreover, the legal and monetary conditions for data access are generally stipulated in the corporation's Terms of Service. It should be noted that Facebook's Terms of Service can be updated at their discretion (with 30 days' notice), and that the company could forbid any ongoing analysis that aims at increasing transparency, simply by changing these terms.

After a first wave of API changes in 2018, Facebook began an extensive review process of the way all its partners used its API in April 2020, which ramped up the following changes in its API in the summer of 2020. Gradually, by way of these changes entering into effect, terms like "social media monitoring" or "social media listening" has come to refer to this type of very restricted media monitor-

ing that customers can apply for. One example of the changing landscape of social media monitoring during this time is how Facebook's Graph Search (a semantic search engine similar to traditional search engines such as Google, introduced in March 2013) was being made less publicly visible as of January 2015, and was almost entirely deprecated in June 2019. Later in 2015, Facebook also depreciated their related API (see below), heralding an era of increased closure and restriction of their data, compared to before.

While research and business communities often refer to the Twitter "firehose" (i.e., APIs offering full-scale access to the data flowing on the platform), there has never been any Facebook or Instagram "firehose," and after the notorious Cambridge Analytica data scandal in 2018–2019 (which made a noticeable impact on the market valuation of the Facebook corporation), access to Facebook user data has become considerably more restricted. In short, the public APIs of these networks need to be accessed through individual user tokens, which are only granted by capacity of having an active Facebook account. The process works by Facebook requiring the users to provide a token to be able to access any Facebook data. By giving the vendor access to one's token, the customer can access a selection of Facebook data streams, of their own choosing: The requirement is that the user actively selects the Instagram and/or Facebook pages they want to monitor.

After this form of verification, the customers can link various Facebook and Instagram pages to their personal dashboard. The vendors' business models are in this respect more about aggregation and packaging of data streams that the customer would already have access to, by capacity of having a Facebook account. Notably, Facebook requires that users demonstrate that they still have an active relationship with the application that is making requests on their behalf. While tokens used to be valid until revoked, many platforms have introduced schemes where their tokens have a default lifespan (measured in days or weeks), and unless users demonstrate that they are still using the application (e.g., by signing on to that application via "Facebook Login"), Facebook will invalidate the token.

The data in question would either be the data from the customer's own Facebook and Instagram pages, or data that results from turning on active monitoring of the activity of Facebook and Instagram business pages that the customers choose to add to their project channels, whether it is to track competitors' social media activities, or to monitor important industry sources or influencers/key opinion leaders. Here, big vendors like Talkwalker offer predefined industry panels external repositories, consisting of aggregated Facebook pages of multiple topics and industries. The data compiled remains exclusive to the customer, and the vendors guarantee that it will not be shared with any other clients.

Customers also have the option to track Instagram hashtags on topics of interest, such as trending hashtags. More recently, vendors have added the option for customers to also include data on the ad performance of their own Facebook and Instagram inventory.

That being said, access to content from Facebook-owned properties is *severely* limited, for commercial vendors of this kind. As we will show in our more hands-on overview of providers and vendors below, the various actors and constellations have been changing rather rapidly. In 2015, before the Cambridge Analytica scandal, Facebook offered, to business users in the US and UK, something they called an "insights product" called "Topic Data" in collaboration with DataSift, a London-based company specializing in brand analytics and data provision. This product was narrowly focused on what audiences are expressing on Facebook about events, brands, subjects, and activities. For the reasons of user privacy stated above, this data was only offered on an aggregated and anonymized basis, so that the business users would not be able to piece together exactly who said what. DataSift stood for a new paradigm in data provision of this kind, in that it did not provide pure API access, it provided already curated feeds from social media data sources, filtered in often very complex ways (High Scalability 2011). Their platform offered dashboards where business users could use pre-designated filters and categorizations and, also, to build upon selections that other users had already made. So, from a user point-of-view, not only was the original data selection and packaging unilaterally defined by Facebook; the preferred partner (DataSift) repurposed this data, defining the terms of access in ways that were often opaque to the end-users, who simply had to put a lot of trust into the service, without being able to scrutinize the actual source data.

Inside Facebook, there is reportedly a tug-of-war between executives who are in favor of more transparency and executives who believe that disclosure of trending articles and topics (e.g., by way of monitoring tools like CrowdTangle; see Roose 2021) might be undesirable from a business perspective, arguing that Facebook should selectively present curated reports of its own choosing or, alternatively, restrict the tools available for outsiders to peruse activities on Facebook (e.g., by only allowing for them to see measures in aggregated and de-personalized form). One important conclusion from this is the understanding that providers like DataSift *would not actually sell the Facebook data*, they merely sold highly curated access to prefab analytics (so-called "insights"), catering to the rather specific needs of advertisers and marketers. This is an analytical observation that is crucial to the often very public and vocal debates that have raged over Facebook, user privacy, and data access ever since.

Facebook depreciated their Graph API version 1 on April 30, 2015. After that, there were no official resellers of raw Facebook data, and the only preferred partner, offering limited access as described above, was DataSift. The only ways to

scrape data, after this date, would have been to either collect highly public data (i.e., data which is from highly publicized profiles, typically brands or celebrities). This data would, however, not be particularly representative of the Facebook public as a whole. Another way would have been to illicitly, or semi-legally act as one or more users within the platform and scrape their activity feeds and posts. A third possibility would be to liaise with existing users, asking for their permission so that data could be scraped pertaining to their (and only their) activities, with these individuals' knowing consent. Importantly, the data collection that led to the Cambridge Analytica scandal some years later had already taken place before these API changes (more specifically, that data was harvested through the third-party app *This Is Your Digital Life*, developed by Cambridge University professor Aleksandr Kogan, that was removed from the Facebook platform in 2015).

In November 2016, Facebook acquired CrowdTangle, an analytics tool that measures how much social-media engagement different posts are getting. This tool is often used by publishers use to track how much traction individual pieces of content are getting but can also be used as an indicator of popularity for different Facebook posts. It forms one element of Facebook's in-house publisher analytics tools, alongside its measurement of "trending topics" and Page Insights (analytics tools for singular Facebook Pages.) CrowdTangle has later become instrumental for assessments of how popularity of content is distributed on the Facebook platform – including a few not so flattering observations, as witnessed by the executives later interviewed by investigative journalist Kevin Roose (2021).

Twitter, being an entirely different social networking platform in that its users are by default public, not private, had opened its API in a much more radical ways a few years prior to this. According to industry insiders, this had "spawned an entire ecosystem of data interpreters, including Adobe Social, Brandwatch, Crimson Hexagon, Socialmetrix and DataSift itself, which was one of only a few companies allowed to sell the full Twitter firehose at one point" (Constine 2015). However, as the competing data provider Gnip was acquired by Twitter in May 2014, it was Gnip that became the preferred partner to Twitter, becoming the only company that provided exclusive access to Twitter data at scale (i.e., not through Twitter's free APIs). By August 2015, Twitter had shifted its entire data-licensing offering to Gnip, robbing DataSift of its access to the Twitter firehose. Since then, Gnip has been integrated into Twitter's business-to-business offering.

In March 2018, DataSift was acquired by Meltwater, a business intelligence company originally out of Norway but primarily based in San Francisco. Significant events during 2018, namely the Cambridge Analytica scandal and the introduction of the GDPR across the EU – meant that business-to-business uses of Facebook and Instagram data have all become exclusively based on modes of access

that do not identify anyone in particular, focusing on aggregated information: "trends," "patterns," and "insights".

> The gist is now any application that is attempting to do work with a platform on behalf of a company needs one or more authorization tokens (e.g., OAuth token, bearer token, etc.) to do that work. This includes common MarTech stack tools such as social media monitoring, listening, analytics, and publishing SaaS solutions. No longer can an application register and make as many calls as it needs to perform the work on behalf of their corporate customer. Access to each API is protected and metered and any access an application is making to a platform needs to be on behalf of (i.e., authorized by) a user. How this access is granted varies by platform. Some platforms are primarily credential-based on an individual account basis (e.g., Twitter), while others are more complex, like Facebook, which have User Accounts, Business Pages, and Business Manager Pages, each with their own access tokens. (Roberson 2020)

Consequently, the current field of corporate partnerships and business-to-business deals in the field of social-media data access is one of byzantine complexity, and our ability to glean insights, as outside observers, is severely restricted by the above setup of strictly monitored, consolidated business partners. What actually circulates? What popularity do certain sentiments have, among what groups of individuals? To an outside onlooker, the field of social media analytics appears bewildering, as a vast range of different companies promise to offer "social media insights" at a monthly cost – but it must be noted that they all operate in a consolidated ecosystem where access flows from the top, dictated by very strict rules enforced by Facebook, as it has evolved from a social networking site to a "platform-as-infrastructure" (Helmond et al. 2019).

2.4.4 Platforms unilaterally restricting academic access

From a social-science perspective, it is increasingly clear that **giant social media corporations (Facebook in particular) employ highly problematic policies for access, transparency, and accountability**. In early August 2021, Facebook suspended the accounts of a group of researchers affiliated to New York University, cutting them off from their on-going empirical work. The team had been working on a project called Cybersecurity for Democracy, aiming at uncovering systemic flaws in the Facebook Ad Library, identifying misinformation in political ads (precipitating distrust in national election systems), and studying Facebook's apparent amplification of partisan misinformation. The researchers gathered data by asking volunteers, with their consent, to install the AdObserver extension into their browsers, in order to help the researchers to collect data. According to Facebook, the extension scraped data that was visible

from volunteers' accounts but contained non-volunteer data that was posted. The company explains that they take unauthorized data scraping very seriously, as it "jeopardizes people's privacy" (Clark 2021).

Interestingly, Facebook, in its above statement, cites an order by the US Federal Trade Commission (FTC). However, in a statement released just a day after it was revealed that Facebook blocked the Ad Observatory, the FTC called this claim "inaccurate". Instead, the FTC said it supports NYU's project and encourages Facebook to exempt good-faith researchers from monolithic and self-serving interpretations of privacy law (Levine 2021).

Later the same month, also another academic research project went public with the admission that it had been forced to shut down its Instagram monitoring project after threats from Facebook. In March 2020, German non-profit organization AlgorithmWatch had launched a research project to monitor Instagram's newsfeed algorithm, a research project supported by the European Data Journalism Network and Dutch foundation SIDN, and in partnership with Mediapart in France, NOS, Groene Amsterdammer and Pointer in the Netherlands, and Süddeutsche Zeitung in Germany. Volunteer users could install a browser add-on that scraped their Instagram newsfeeds. Data was sent to a database we used to study how Instagram prioritizes pictures and videos in a user's timeline.

The Facebook terms-of-service stipulates that one "may not access or collect data from [Facebook's products] using automated means" (Kayser-Bril 2021). However, AlgorithmWatch argue that they only collected data related to content that Facebook displayed to the volunteers who installed the add-on. "In other words, users of the plug-in where only accessing their own feed, and sharing it with us for research purposes." Facebook also claimed that the system violated the GDPR "because some of the collected data stemmed from users who never agreed to the project, whose pictures were shown in the timeline of our volunteers." However, AlgorithmWatch argue that their open-source code reveals that such data was deleted immediately when arriving at their server, and that the add-on was deliberately designed with great care so that Facebook would not be able to identify and prosecute the volunteers. In July 2021, after significant pressure from Facebook, AlgorithmWatch decided to terminate the project and delete all collected data (letting media partners still have fully anonymized versions of the data, however).

Not only the FTC has expressed concerns. In 2019, the US Social Science Research Council (SSRC) launched a program called *Social Media and Democracy Research Grants*, in collaboration with Facebook, with the explicit purpose to "make privacy-protected data available to social researchers to examine Facebook's impact on elections and democracy" (Nelson 2019), as part of the larger industry-academic partnership hub Social Science One (see King & Persily 2020 for

more detail). However, it soon transpired that Facebook had, according to the official statement, underestimated the complexity of the task of delivering legally workable data for research purposes, and was unable to deliver all the data initially anticipated. The project was subsequently discontinued.

Consequently, in December 2019, the European Advisory Committee for Social Science One (hosted by Harvard's Institute for Quantitative Social Science) had to put out a statement, signed by a group of leading academic scholars led by Claes de Vreese (University of Amsterdam). They noted that Facebook had, after 18 months of negotiation, still not provided them "with anything approaching adequate data access" (Social Science One 2019). Because Facebook had not been able to provide even the initial, aggregated dataset of URLs shared on the platform, Social Science One's philanthropic funders had begun to withdraw:

> As members of the European Advisory Committee of Social Science One we – along with the co-chairs – are frustrated. On the one hand, we were genuinely interested in helping to build a model to support academic research, and we appreciate the efforts of the specific data science teams within Facebook have made to this end. On the other hand, the eternal delays and barriers from both within and beyond the company lead us to doubt whether substantial progress can be made, at least under the current model. The current situation is untenable. Heated public and political discussions are waged over the role and responsibilities of platforms in today's societies, and yet researchers cannot make fully informed contributions to these discussions. We are mostly left in the dark, lacking appropriate data to assess potential risks and benefits. This is not an acceptable situation for scientific knowledge. It is not an acceptable situation for our societies. (Social Science One 2019)

More recently, it has transpired that even those projects that were funded by Social Science One and made use of Facebook data might have been compromised, since data that researchers were given by Facebook had referred to populations that turned out to be erroneously defined; for projects that aimed to chart engagement with political content, e.g., the data had been claimed to involve all US American Facebook users, but later it was discovered that the data actually included the interactions of only about half of Facebook's US American users – namely, the ones who had engaged with political pages enough to make their political leanings clear (Alba 2021). This indicates that Facebook, first, had been in the position to make claims about that data which were hard for the collaborating researchers to verify and, second, that Facebook as a corporation might not have cared much at all about the actual reliability of that data.

Another group of researchers have explored how the data masking, a method that Facebook employed for its large datasets available to researchers, seems to have affected the representativity of the datasets in question. For an enormous

dataset that Allen et al. (2021) had perused, containing 10 trillion cells, and consisting of URLs shared on its platform and their engagement metrics, Facebook had deliberately done two things to protect the privacy of its individual users. Needless to say, the data did not contain any information on individual accounts to begin with, merely the aggregated counts of shares for different public URLs on the internet. Moreover, Facebook had (1) added differentially private noise to engagement counts, so that it couldn't be inferred from the counts what actual accounts the aggregated counts referred to, and (2) censored the data so that all URLs with less than 100 public shares had been omitted from the dataset to begin with. However, Allen et al. (2021) found that this latter method affected the representativity of data in a negative way. When they compared the distribution of fake news in the massive, censored URLs dataset with an estimate from a smaller, representative dataset, they found that the censored dataset overestimated the share of fake news and news overall by as much as four times:

> Content that is likely to be shared conditional on being viewed or clicked, and conditional on being shared, be shared publicly, is overrepresented in the Facebook URLs dataset. Some content, like fake news, is optimized both to be clicked, since it is likely to be novel and have click-bait headlines, and to be shared publicly, since it is likely meant to draw others to engage. On the other hand, retail ads are optimized to be clicked due to personalized targeting, but not to be shared publicly, since they might contain private information that users would not want to disseminate. For example, micro-targeted Facebook ads containing links to political fundraising would likely be underrepresented in the Facebook URLs dataset, despite being of interest to researchers, because those types of URLs are unlikely to be shared publicly. (Allen et al. 2021)

In other words, deliberate omissions or editing of datasets should be done with caution, exploring the actual repercussions of such curatorial practices, something that we will revisit below, in the section on data quality of Swedish-language URLs (3.3.1). It is problematic if specific societal actors have the power to unilaterally make such decisions without any oversight or ability for other societal actors to verify the quality of information disclosed.

At the same time, there are numerous academic-industry collaboration agreements involving legal setups whereby researchers are bound by contractual limits to potentially confidential or harmful information. Take, e.g., the UK Lambert Toolkit (Intellectual Property Office 2019), intended to help facilitate negotiations and secure agreements between academic or research institutions and industrial partners. This has been hosted and approved by the UK government for almost 20 years and is used rather extensively in the UK for academic-industry collaborations. Nevertheless, if a company can unilaterally define the terms for which protecting "confidential information" would apply, this would amount to a skewed power relationship vis-à-vis the academic research community – and if

that company wields such momentous market power as Facebook does, this is indeed hugely problematic, since the market in question is not just any market, but the very *agora* for democratic deliberation among civic actors in societies, worldwide.

It should be added that Facebook is not the only platform that is characterized by secrecy, lack of transparency, and omnipotent agency to control the terms of access to its data; arguably, successful audiovisual media platforms like TikTok and YouTube, e.g., are even less transparent, regarding what circulates and how popularity is distributed.

2.4.5 The challenges of genre categorization

Analytically, **the categorization between "editorial/news" on the one hand, and "social" on the other is sometimes spurious and rather arbitrary**. Consider a blogger who regularly publishes posts on a blog platform, mainly consisting of quotes from editorial news stories, and very little else. Is this blog to be considered a publicly facing text-based outlet for a private citizen, or is this blogger effectively acting in an editorial capacity, when selecting and repurposing, even directly re-distributing access to professional news outlets via links on his/her blog? If this private citizen happens to make some money from this blog, when is it to be considered more like an actual editorial outlet?

The internet is beset with problems that has to do with implementations of tasks like sorting and categorization, *at scale*. Often, service providers like online text harvesters make decisions that generate "path dependencies" as a particular website or list of websites is tagged according to a set of definitions, after which the algorithmic scraping and sorting does its job and keeps tagging these websites accordingly. But the actual websites might change purpose, form, and direction over time, while the data harvesting algorithms make no notice of such changes. Thereby, sites that might originally realistically have been tagged "social" might over time have changed into much more automated, nonhuman indexing sites that merely scrape the web and re-post existing news stories. So-called "news aggregators" are very common in datasets of online text.

How to categorize blogs that seem to have editorial ambitions? In some countries, the publishing legislation requires a formally registered publisher, however it is not clear as to whether there is commensurable legislation in every country. Typical criteria for the operational definition of "published" texts are given by Biber (1993): "(1) they are printed in multiple copies for distribution; (2) they are copyright registered or recorded by a major indexing service. In the

United States, a record of all copyright registered books and periodicals is available at the Library of Congress" (p. 245). If we are to follow that definition, many bloggers – regardless of how ambitious and scrutinizing they are – fail to qualify as publishers.

How to categorize news aggregators? These merely re-distribute already published news stories, and aggregators comprise a nontrivial part of online editorial text data. Sometimes there are hybrid categories, also in that there are news aggregator sites that allow for social commentary on each individual news story.

Language hybridity: When verifying websites, several individual websites appear to be English based. Yet, they might contain singular posts other languages; in these cases, what to look for is that the ISO tagging of individual posts is correct, in terms of this tagging corresponding with the actual language of the post. For multilingual sites it cannot always be guaranteed that each individual posts would be correctly tagged; see for example Malaysian multilingual editorial news site *bernama.com*.

3. Data provision in context

3.1 An overview of the existing ecology of data providers

First, in this report we will **differentiate between the broader term "provider" and the narrower term "vendor"**. In the OECD's glossary of terms (OECD 2013) a *data provider* is any organization which produces data or metadata. This should not be confused with the legal terminology in the EU's General Data Protection Regulation (GDPR) (EU 2016/679), which specifies "data controllers" and "data processors." If the data in question pertains to EU citizens, both the provider and its database users are subject to GDPR compliance. In order to send any EU citizen data to any Data Provider for any purpose, a signed Data Processing Agreement (DPA) or equivalent contract must be produced between the provider and its customers. Note that GDPR does not just apply to EU-based companies, but to any company worldwide that holds EU citizen data.

> The duties of the processor towards the controller must be specified in a contract or another legal act. For example, the contract must indicate what happens to the personal data once the contract is terminated. A typical activity of processors is offering IT solutions, including cloud storage. The data processor may only sub-contract a part of its task to another processor or appoint a joint processor when it has received prior written authorization from the data controller. There are situations where an entity can be a data controller, or a data processor, or both. (European Commission, n.d.)

A *data vendor*, in this context, should be understood as an entity that provides access to data on a commercial basis. While a data provider could be any entity, also nonprofit or government-funded ones (CommonCrawl is one such example, described in more detail below), a vendor is typically a company operating in sectors like market intelligence, social media monitoring and/or listening, digital advertising, and the like.

Using a data provider to access online text data has its benefits and its drawbacks. Needless to say, a project like LES has had very specific **data requirements**. In order to build all models that make up the Lexicon, the project members needed to access text data in multiple languages from multiple countries and also needed to know what type of media the text sources came from, since different media types contain different forms of text content. On top of these specificities, vast amounts of data were needed to be able to train each specific model. Collecting so much data is normally extremely taxing, unless one has the financial and technical resources of a larger provider.

In the LES project, many of the researchers involved have been affiliated with the Swedish language technology company **Gavagai**. Since the launch of Gavagai in 2008, some of the researchers involved in the LES project have also

been instrumental in the setup of this research-driven company, which explicitly focuses on real-world NLP applications (using unstructured language/text data and making automated inferences from such data through different machine learning techniques). A dataset sourced from a range of different commercial providers was organized through Gavagai and proved highly useful, primarily because this provided enough data and allowed the project members to build models from a very large number of languages and countries, while retaining a preferable variety and volume of data. This data was deliberately curated to improve the frequency of documents also from smaller languages. Moreover, the LES project members also had the ability to suggest to the provider various source URLs to include in order to maximize quality and minimize potential noise. That being said, there are still various challenges when working with either commercially provided data or when scraping data from the Web by brute-force (e.g., CommonCrawl).

Language **data size variability** is problematic, as scarcity of data (a problem that pertains to many countries and languages in Africa and Asia in particular) can prevent analysts from using several of the existing models and/or from building new models for more languages and countries. The primary reason for this scarcity is that many poorer countries simply have a very modest supply side of online news and user-generated media, in comparison to many richer countries, something that will be apparent from our selection of country cases below. But it's also related to the commercial nature of the data; a vendor has an intermittent interest in crawling large websites in, e.g., Somali from Somalia unless their customers specifically have employed them to do so.

It is important to note that in the LES project, the specific data requirements of the project – in particular, the need for multiple data splits in terms of languages and sources – were instrumental to the choice of data provider. While there are non-commercial data services (e.g., CommonCrawl) that can provide open text data from billions of different websites across the world, such services do not guarantee the specific data splits required to build all the models currently covered by the Lexicon. Moreover, using existing crawl data requires quite intensive processing, as each release contains billions of freely available web pages from a range of locations and languages. The data was not cherry-picked due to commercial interests but, equally, the process of curating text data to exact specifications required (like, e.g., when building a lexicon) would in itself be an overwhelming task.

Language data from a commercial provider can, for example, be **geo-located and/or source-tagged**, and, moreover, guaranteed by the provider to be relevant to the initial categorization (social/editorial) which is hugely advantageous for purposes like the ones that the LES project specified. However, in using a vendor, one is also forced to accept the realities of a commercial landscape. This

effectively means that one would have very limited control over which data is originally collected, since this is largely determined by the needs of the vendor that in turn caters to the needs of its customers. There are some critical consequences of using commercial data, particularly regarding data representativity and replicability.

In terms of the text data content, we would like to emphasize that online data – and social data especially – is a **lucrative commodity**, with its own specifications and requirements on the demand side. Procuring data from large technology companies is difficult and very costly, even for larger data distributors. Sometimes, Web documents located behind paywalls are omitted (due to contractual terms), and so would also data from closed platforms such as Facebook or Twitter be. Moreover, certain types of data are very rare to come by (private conversations in particular), and historical data (older than 30 days) is rarely made available. As discussed above, such platforms, while not necessarily being directly representative of a general population, are nonetheless widely used across the world, and their exclusion from the text corpora may consequently result in loss of valuable user-generated content. In sections 2.4.3 and 2.4.4, we have detailed how the current oligopoly of social media platforms severely restricts the conditions of access for both commercial providers and nonprofit/academic researchers.

Equally important, Web documents tend to be **classified beforehand** by the data vendor, and as customers to the vendor's services, one would have little insight into the criteria under which all the numerous data sources are being classified, and if the criteria are the same across all languages and countries. In view of the ever-changing landscape of the internet, defining *source types* – news media, social media, or forum media – is particularly challenging. Today, news articles are often accompanied by comment sections where parts of the reader base continuously create content that cannot be defined as pure news media, and it remains unclear as to how such sections are being classified. The rise of alternative media constitutes another challenge, not least because such websites largely contain editorial content and are often made to look like those of regular news media. Moreover, the growing popularity of online marketplaces and other listing platforms further complicates the matter. Platforms like TripAdvisor or Yelp have a vast user base that in turn generate a substantial amount of text documents, yet they cannot strictly be considered as social media. From this perspective, the labels 'news' versus 'social' media is sometimes a bit misleading and a more appropriate notation would perhaps be 'legacy versus 'new' media, where 'legacy' refers to traditional newspapers (albeit in digital form) and 'new' media to everything else.

Whilst it is tempting to call for more transparency into the commercial practices of collecting and distributing online data, it is important to remember that **copyright and strengthened legal frameworks for personal data protection**

and privacy – such as the GDPR – prevents many companies from disclosing such information. In effect, such data often comes with the proviso that sharing the raw text data (or even mere access to it) with third parties is restricted, which diminishes the possibilities for good replicability of the methods used (e.g., the exact ML model outputs).

It is not our intention to provide the ultimate list of what vendors to use and their current offerings, since this is a moving target. Vendors get bought and sold all the time, they come to include new services into their offerings, or change their terms and conditions or technical solutions for data access, and they even see some of their data sources being redacted, like with the Facebook corporation's different properties. Rather, what we are trying to present here is a tentative overview of key aspects to look for, when orienting oneself in the field of semantic data provision, as well as reasonable critical questions to ask, when assessing the suitability of various vendors' offerings to serious academic research projects. Hopefully, this guide will help the aspiring researcher in knowing what to look for, and what to ask for when negotiating with vendors and providers.

3.1.1 A tentative list of some data providers

The supply of commercial providers of editorial and user-generated text data, for different languages and in different countries, forms a complex ecosystem. Access to user-generated online text data is determined by the largely commercial nature of not only the sources, but also the redistributors of such data; complex arrangements of interlinking commercial vendors, each providing different modes of access and collections of sources (generally by way of different APIs and/or visual dashboards).

Each vendor offers a particular selection of sources, and moreover remains largely opaque as regards the provenance of these sources. Some data types, as we have seen, are very rare to come by. Facebook data isn't available as raw data (even for commercial purposes) but only as prefab analytics, with access terms unilaterally controlled by Facebook, and, generally, broad access to Twitter data is also hard to come by (also Twitter has moved to a model where access to the firehose goes exclusively through Twitter, with only more restricted modes of access and terms of use for third-part providers). Rather, the data in question is often scraped from wikis and other online corpora, blogs, and message boards on the open Web, alongside various Web-based editorial news sources. Moreover, commercial vendors normally only make available the most recent thirty days of data, shelving older data on magnetic tape (thus making it significantly less accessible). Each vendor also presents its own, institutional terms of use, by default often making re-use of the data in question impossible. For these reasons, and

several more, it is hard to employ traditional standards of statistical representativity and replicability. In short, certain types of data, in particular, are generally very hard to get hold of:

- **Chat app data from, for example, WhatsApp, WeChat, Snapchat, Kik, and Facebook Messenger.** This is, by default, extremely private data and, as such, highly sensitive.
- **Individual posts (private and public) from Facebook, Instagram, and LinkedIn.** Not perhaps equally as sensitive as the above data, most of these posts are by default private (note that the in early instantiations of the Facebook platform, the default setting was not nearly as private as it soon became, in later instantiations).
- **Geo-tagged data.** There are ethical and legal reasons for users not consenting to having their geographical position tracked; moreover, there are technical reasons as for why this can be hard to reliably attain.
- **Demographic variables (e.g., gender, age, and income) for the sources in question.** Like with the above, this is both sensitive out of privacy concerns but also technically challenging to attain, unless users were forced to mandatory registration of their identities online.
- **Historical data (older than 30 days).** Providers rarely keep older data available, at scale, since such provision is costly and makes little business sense in a market where the demand side tends to value immediacy and topicality.

For social media monitoring sites to have access to historical data, they would either have to keep in-house historical archives or buy this service from someone who does. APIs tend to only go back a couple of weeks and Facebook doesn't offer a historical search API going back further. It is very much the same for Twitter and other social media platforms. Many social media monitoring tools, however, maintain an archive of data going back a year or more in time which they can query directly. Companies like Gnip and DataSift also offer historical data query services which can be used to go back further. Gnip claim to have every tweet ever made; Datasift only has tweets back to 2010, but it does also offer historical data for other sites such as Tumblr, Facebook and Bit.ly.

Some interim conclusions can be noted. To begin with, few vendors provide adequate **service-level agreements concerning data coverage and/or latency**. There are generally no formal guarantees whatsoever concerning **text quality**. Vendors tend to guarantee only certain quantities of text and/or frequencies of specimens.

As has already been noted, **legal conditions** – such as copyright, data protection (e.g., the GDPR), and terms of service – dictate a lot of archival uses of web-medi-

ated data. These conditions are stacked on top of each other: The original platform (e.g., Twitter) would stipulate certain terms of use, but also the attendant commercial retailers stipulate their own, extraneous terms of use, often explicitly prohibiting any sharing of the data, due to its business value.

All these factors are inimical not only to the provenance, reliability, and representativity of the data—but also to **reproducibility** of scientific results. Despite searching for the same time period and language, it is not hard to imagine one vendor providing a slightly different set of data than another vendor, depending on how they handle things like geo-tagging and/or language detection, original API access, search query designs, and so on. Even the point in time when the query is made would have effect, since the original platform might have removed old content at any time. There are many potential points of failure and/or bias, and providers are unlikely to give any formal guarantees regarding data completeness.

To enable at least formal reproducibility/replicability of scientific results, one can employ large-scale Web scraping services, for example, CommonCrawl, or commercial tools for scraping (so-called "crawler-as-a-service" tools), as alternative means to gather data for the purpose of large-scale processing and pattern recognition. However, **manual Web scraping is nontrivial**: Data is always noisy and requires considerable processing before use. For any scientific research project, this suggests that considerable resources should be considered for the purpose of scraping, cleaning, and preparing data alone.

There is an **almost endless list of companies** operating in the broader fields of business intelligence and media intelligence, often specializing in media search, media monitoring, and media analytics (both editorial and social). New market entrants crop up almost every week, in different legislations around the world, often focusing on specific national or regional markets. Some of the providers of social media monitoring dashboards (e.g., Hootsuite, CrowdTangle, Quintly, Socialbakers etc.) do provide intelligence based on cumulative data from their own properties and can therefore assess things like popularity metrics for specific social media in specific countries. However, as we will note below, only a small segment of these companies actually offer linguistic data that is more comprehensively representative of larger populations – i.e., offering monitoring of the surrounding, online public realm beyond the much narrower focus on marketing metrics for individual clients' own online properties.

Since the market is largely dominated by recently founded companies, primarily financed through venture capital (private equity), and therefore not publicly listed, **transparency is lacking** and very much of the total market remains opaque to outside observers. Hence, what is offered below is a glimpse of the total market, with numerous brands and names left out for brevity.

We identify a few actors that are placed higher up in the **value chain**, offering services that can be likened to 'bulk' data, refined by subsequent services further down the value chain. A handful of such source providers identified are Gnip (USA), Socialgist (USA), Opoint (Norway/Denmark), and Webhose.io (Israel). A very significant non-profit provider to consider alongside these commercial actors is CommonCrawl (USA). As for CrowdTangle and DataSift, while claiming to offer unique insights into Facebook's data stream, both of these companies suffer from Facebook's infamous shift in policy towards greater opacity, as has recently been shown in investigative journalism: In April 2021, CrowdTangle "which had been running quasi-independently inside Facebook since being acquired in 2016, was being moved under the social network's integrity team, the group trying to rid the platform of misinformation and hate speech" (Roose 2021), since the tool (despite its limitations) was deemed to enable glimpses into the distribution of popularity of Facebook Posts that were unfavorable to the public perception of Facebook as a brand. A structural reason for this conundrum is that only 'engagement' data are made available by Facebook, not 'reach' data (except for video views). The latter, it could be argued, ought to be a potentially more comprehensive and reliable measure of actual popularity, since the former metric ('engagement') is often particularly high for those types of content that have a sensationalist and politically polarizing slant (see the discussions about "media logic" above).

Arguably, a degree of **competition and structural differentiation** between actors of these kinds is beneficial to the provision of more comprehensive, reliable, and less biased data: If platform giants like Facebook had their way, only such data would be made available that wouldn't risk making the platforms appear ethically problematic or less popular than they would prefer to be. Market dominance of this kind seems incompatible with proper transparency. It seems reasonable that strong and independent third-part data providers can be able to exert pressures on platform giants to provide more comprehensive, useable data.

Two companies that are to be considered more as holding companies (consortia) than as providers *per se* are Meltwater (owning Sysomos, DataSift) and Infomedia (owning Opoint, strategically collaborating with Talkwalker, Hootsuite, and ZeroFox).

Regarding the provision of purely editorial text (i.e., sourced from the press or other legacy media), we observe a handful of companies that inhabit a similar role in the value chain as the abovementioned social media data providers. Several of these news-oriented companies have long-running legacies of conducting in-house, proprietary news scraping and archiving. Some key examples are, on the transnational market (e.g., LexisNexis, Bloomberg, Thomson Reuters, etc.) on and local/regional markets (e.g., Retriever in the Nordic region).

Central actors (in no particular order)

CommonCrawl. California-based non-profit organization that makes monthly crawls of the openly available Web and provides datasets and metadata to the public freely. The CommonCrawl corpus contains petabytes of data including raw web page data, metadata data and text data collected since 2011. Since 2012, CommonCrawl's archive is hosted by Amazon Web Services as part of its Public Data Sets program. Every crawl contains around 300 terabytes of data and roughly 3 billion pages. In 2020, a filtered version of this CommonCrawl archive was used to train OpenAI's GPT-3 language model.

CrowdTangle. California-based media monitoring service, introduced in 2011. Originally designed as a tool to help communications specialists organize their Facebook activity through a single dashboard, with the added feature of measuring engagement rates across different Facebook Pages. Attracted private investment funding, only to be acquired by the Facebook corporation for an undisclosed sum in November 2016.

Gnip. Colorado-based data provider, founded in 2008 and more operationally functional by 2010, focusing on social media API aggregation. After a 2010 data licensing agreement with Twitter, Gnip was acquired by Twitter in May 2014. By August 2015, Twitter had shifted its entire data-licensing offering to Gnip and integrated it into Twitter's overall business-to-business offering.

Datasift. London-based company, founded in Reading (UK) in 2010, specializing in brand analytics and data provision. Now operating primarily out of San Francisco. Over the course of the early 2010s, DataSift developed partnerships with several tech companies including Twitter, Facebook, LinkedIn, WordPress; the partnership with Twitter ended in 2015, as Gnip became Twitter's exclusive data partner (DataSift's Twitter firehose API contract was terminated in April 2015, access was finally switched off in August the same year, as Twitter's new firehose APIs were released in October), while DataSift offered a customised product based on an exclusive contract with Facebook. Following the Cambridge Analytics scandal, Facebook has become even more restrictive with its data properties, and it's rather unclear as to what status DataSift has, in terms of access to Facebook analytics, compared to its competitors. DataSift was acquired by Norwegian company Meltwater in March 2018.

Datasift offers LinkedIn Engagement Insights – aggregated actions from millions of users, categorised across 130 attributes — a similar product as the (now-depreciated) Facebook "Topic Data." The company also offers firehose access to WordPress, Tumblr, Disqus, and news articles from Lexis Nexis and NewsCred.

Meltwater. One of the early entrants to the field, Meltwater was founded in 2001 by Norwegian businesspeople, but is based in San Francisco since 2005. Specialising in media intelligence and social analytics, the focus is not only on social media monitoring, but, perhaps even more importantly, editorial text sourced from select publisher partners; like some of its competitors at the turn of the millennium, Meltwater started as a news clipping service. It is a privately held company, however currently listed on the Euronext market index.

Sysomos. Toronto-based social media monitoring and analytics company, founded in 2007. Its initial focus was on blog monitoring and geolocation-based social search (Foursquare). Operated as a subsidiary owned by Marketwire between 2010 and 2015, as Sysomos went independent in February that year. Three years later, Sysomos was acquired by Norwegian company Meltwater (April 2018), continuing to operate as an independent unit within their consortium.

Opoint. Norwegian media monitoring and analysis company, founded in 1996. Developed a proprietary web-crawling service and operating several similar web-crawling operations in different languages and markets. Currently offering a structured data feed with content from more than 170,000 websites worldwide, alongside 20 years of structured historical news data. In 2016, Opoint was acquired by Finnish information services, technology, and advisory company M-Brain, only to be sold two years later to Copenhagen-based Nordic media intelligence company Infomedia A/S.

Infomedia. Danish media intelligence company, established in 2002, owned by two of the leading Danish media corporations, JP/Politikens Hus and Berlingske Media. Offering an array of services, mainly in the fields of media monitoring, media analytics, research, consulting, etc.

Infomedia owns Opoint, which acts as a provider to Infomedia's suite of products. As of March 2018, in the Nordics, Infomedia acts to implement the products of partner companies Talkwalker, ZeroFox, and Hootsuite. The value added from Infomedia, in this collaboration, is its consulting team, offering expert advice for the implementation and analysis.

Talkwalker. Founded in Luxembourg in 2009 under the name Trendiction, focusing on crawling online editorial media and social media. The name Talkwalker was adopted through a merger in 2016. Opened its first overseas office in New York in 2015 (later also San Francisco and Frankfurt).

In the context of this report, Talkwalker is one of the more interesting providers, as it offers a vast range of languages; at the time of writing, they claim to offer

social and editorial monitoring in 187 languages. Its business offering is very dependent on its browser-based dashboards, visualizing various metrics, powered by an in-house AI engine, offering text sentiment analysis and automated image recognition. Monitoring and analytics appear to be their primary business, not provision of raw data.

Regarding Talkwalker's social-media data collection, their initial attempts were based on the company's own, proprietary Web crawling. Over time, this sourcing has been complemented by data provision from Socialgist[6] and, most likely, Opoint. Talkwalker is likely to get its editorial data from Opoint. Talkwalker claims to index over 40 million documents per day, categorizing and analyzing these by means of automated data enrichment processes. Their 150 million different sources include news, blogs, discussion boards, online forums, etc.

Webhose (subsequently, **Webz.io**). Data provider based out of Tel Aviv, Israel, founded in 2016. The company describes itself as a "web data provider turning unstructured web content into machine-readable data feeds" (Webhose, n.d.), providing a data-crawling API solution that downloads, cleans, structures, and organizes raw data into datasets its users can consume. During our period of research (in 2021), Webhose claimed to be covering 76 languages in 230 countries, with 10 million posts indexed per day from one million different websites. The company's primary sources are news, blogs, forums, reviews, e-commerce, dark web pages, and some broadcasting (US-based sources). By means of automated data processes, they enrich their data by way of identifying named entities, sentiment, categories, and countries. In late 2021, the company was re-named Webz.io.

Two standout features are API access (providing a "News" API and a "Blogs" API) and access to stored historical news articles since December 2014, some data even dating back to 2008. Like Twingly, Webhose has been of interest to the LES project, thanks to its comprehensive regional coverage, striving to represent a broad range of languages and countries in its offering.

Twingly. Twingly was founded in Linköping, Sweden, in 2006. Like Sysomos, their data mining initially focused on blog indexing. Further, they saw a leading edge in providing coverage of European languages, hence having an offering that is being comparatively more regionally focused than many of the US-American competitors, categorizing and analyzing the data through automated data enrichment processes, employing several categorizations in 35 languages, with also entity and sentiment recognition for some of these languages. Like Webhose, Twingly provides API access, indexing blogs, forums, news, dark web pages, and also Russian social media posts on VKontakte.

---

[6] This has been verified by sources (unnamed Socialgist executive).

Twingly claims to make available over a million blog posts added per day, 8,000 new active blogs added every day. 10 million forum posts added per day, from 9,000 different online forums. 3 million news stories per day; 150,000 sources; over 100 countries. Twingly used to scrape public Facebook pages, up towards 17 million posts per day, but make no mention of it anymore.

Socialgist. Based in Troy, Michigan (USA), founded in 2000. Socialgist brands itself as more raw-data oriented, as it also serves as upstream provider to many of the more user-oriented dashboard services mentioned in this report (e.g., Sysomos, Talkwalker, etc.), seeking to identify, index and make social data available in a structured way, using proprietary, in-house infrastructure and datacenters. Acquired by leading Japanese social data provider and media analytics firm Hottolink in 2014.

The company offers both search APIs and streaming APIs, with custom-built data feeds and integrations. Indeed, they do not even provide a graphic interface, only the raw APIs. One leading edge of Socialgist is that they offer access also to Chinese-language social media (Sina Weibo, Tencent), Russian (VKontakte), Reddit, and, more recently, Quora and Tumblr.

LexisNexis. As a corporation, data mining stalwart LexisNexis covers many more markets than are covered in this report, operating in legal, risk management, corporate, government, accounting, and academic markets. The New York-based company was founded in 1970 (before the personal-computer era) specializing in provisioning electronic access to legal and journalistic documents. Soon, the company branched also into provisioning of archival records, both academic and news/editorial. Besides the company's more recent focus on financial intelligence (risk management, fraud detection, etc.), the name hints at its twofold operations: Lexis – providing access to legal databases and public records; Nexis – providing access to news and business sources, both legacy (print) and Web-based.

Retriever. Leading Nordic provider of archival news media text, media monitoring and analysis, founded in 2002 through a merger between Nordiska Nyheter (founded in 1999) and Infobilis (founded in 2001). Acquired in 2004 by Schibsted, who later divested Retriever and sold it to Norwegian news agency NTB and Swedish news agency TT Nyhetsbyrån in August 2009. Runs proprietary news archive Mediearkivet, alongside a social media monitoring and analysis dashboard. By capacity of doing in-house news scraping and being a subsidiary of the leading Nordic news media companies, Retriever is the regional market leader when it comes to news text.

Quintly, Socialbakers, Radian6, etc. As we have already noted, there are numerous companies (e.g., Quintly, a Germany-based social media analytics company founded in 2011; Socialbakers, a similar Czech company founded in 2008; Radian6, a Canadian social media monitoring platform founded in 2006), that lay claims to providing comprehensive metrics for all (or at least most of) the large-name social media platforms. However, for Facebook, Instagram, Twitter, Snapchat, LinkedIn, Youtube, and even TikTok, the data in question tend to be the client's own performance metrics from the client's own inventories on these platforms. Fernando van der Vlist and Anne Helmond (2021) have mapped, in excruciating detail, the "exceptionally complex global and interconnected marketplace of intermediaries involved in the creation, commodification, analysis, and circulation of data audiences for purposes including but not limited to digital advertising and marketing," in which the majority of these business-to-business-oriented actors are either 'data marketplaces'/'data providers' (e.g. data brokers, suppliers, vendors) or specialized in data analytics and advertising technology ('adtech') – or both. We would go so far to state, however, for the purposes of this report, that very few of these actors actually engage in scientifically useful data of the kind that we are interested in, since the key purpose of almost all the actors listed by van der Vlist & Helmond (2021) is "to map digital traces onto individuals" – i.e., not providing a reliably objective, disinterested snapshot of what discourses actually circulate in social networks and on Web-based forums. The primary aim of these companies is to offer products and services that help clients keep track of the clients' own marketing vectors. Based on such marketing data, these companies tend to offer their own custom dashboards, categorizations, metrics, report templates, APIs, and so forth. The data in question, in other words, is nothing like the semantically rich, ideally representative public data that this report seeks to explore.

3.1.2 Ethical and legal issues

Privacy, transparency, and intellectual property are core ethical concerns here, and often turn out to be what restricts researchers and analysts from being able to access to or work with specific types of data. Below, we will address a few of the challenges that we could identify when working with data from providers, both commercial and non-commercial.

To begin with, there are several parallel legal vectors that must be considered:

- **national legal provisions** (which are, in turn, often produced as part of binding transnational agreements, e.g., the GDPR regulations in EU member states)

- specific **rules and regulations for the specific area of practice or implementation** (for academic research, this includes formalized structures for ethical review that are normally a requirement for funding)
- more specific **licensing agreements** pertaining to the actors involved (e.g., terms and conditions for using specific platforms and thereby accessing their data)

Dahlberg et al. (2021a) have explored GDPR issues for social science, and survey methodology in particular. They have interviewed researchers and policy specialists in the Nordic region and find that when it comes to the ongoing work of individual researchers, these researchers were not significantly affected by the implementation of GDPR – even though the new legal framework has led to stricter requirements for, for example, information for possible participants in research projects and that awareness of privacy issues has increased among the general public, researchers, data collectors and relevant authorities. The difficulties that do crop up – and that may, at first glance, be thought to be due to GDPR appear, on closer inspection, to instead be challenges that arise for other reasons such as technology and commercial licensing, ethical review issues, or differences between different national jurisdictions in areas such as publicity and secrecy (p. 24).

One key problem that has emerged is to do with the ambiguities about what counts as *personal data*, which in turn has affected both data collection, research collaborations, and the publication of research data. Despite the definitions stipulated in the legal text, the term "personal data" sometimes allows for different interpretations. The meaning of the term may vary, depending on context. There is general concern that universities, since they are often obliged to preserve and archive public documents due to "freedom of information" clauses (at least in Northern countries), this obligation might clash with the general stipulation in the GDPR that personal data are to be discarded when no longer needed. A consequence of this, Dahlberg et al. point out, may be that universities can no longer conduct survey research as they have usually been able to, and that academics become relegated to the private sphere for their data collection (p. 24–25).

From a legal point-of-view, one initial parameter to consider when dealing with social data access, is whether the data in question is "crawled data" or "licensed data."

Crawled data is garnered through, essentially, trawling and scraping the public-facing Web, much like Google indexes Web pages through its search index. One such example of crawled data is how, in January 2021, German social-media monitoring company Quintly proclaimed that they are now offering TikTok analytics. This is something of an endeavor, given that tracking this ever-changing

mobile, audiovisual social network "at scale and without any public API or support from TikTok" (Grzesiek 2021) must be daunting – not least from a legal point of view. Quintly apparently manages to scrape TikTok data by scraping the publicly accessible data that is also visible and indexable by search engines.

For manual scraping, numerous approaches exist within the academic community, often requiring workarounds and kludges and some proficiency in coding languages like R and Python. Researchers have argued in favor of Reddit, for example, as it is both "both targeted and free," and moreover demonstrating validity, reliability, and greater demographic diversity than student samples (Jamnik & Lane 2017). Additionally, online forums like Reddit constitute places to recruit live respondents as well. Several tools have been developed for automating data collection from Reddit, e.g., the RedditExtractoR package (Rivera 2019) for the free and open-access statistical software R. Likewise, Twitter is commonly used as a data source, thanks to its relative openness and the public nature of its discourse. While Twitter as a company provides ample documentation for academic researchers on its Developer Platform, as we will see below it's important to note, however, that the permissions that Twitter gives to researchers using its platform must be compliant with what is demonstrated in Twitter's provision of a Developer API and the terms that go with it, as Nicolas Gold (computer science researcher, UCL) notes in a recent ethics overview (Gold 2020). "In essence, Twitter offers its platform for research (but only under certain conditions). Note that scraping Twitter is not permitted as a method to access its data" (p 5). Those who want to scrape data (i.e., researchers) "must inform Twitter of their intentions at the outset (and if these change) in order that it can approve the proposed work. As such, Twitter imposes an informed consent process that meets its own acceptability criteria. Nothing more than active compliance with the terms would therefore be required from an ethics standpoint (unless a researcher intended not to comply)" (Gold 2020: 6).

Licensed data, conversely, refers to the data that a provider would get through entering an agreement with the data owner, normally a platform (e.g., Twitter, Reddit, LinkedIn, etc.) The terms of the license dictate all the details regarding how the data is to be delivered, controlled, and maintained.

Data provider Socialgist has, for example, recently announced two new strategic partnerships, with the aim of being able to provide licensed instead of crawled data from two big American platforms; Reddit and Quora.

The advantages of licensed data are that it's generally more quickly assembled (in real-time, or near real-time), more well-structured (hence more actionable), as well as safer and less risky to handle (since it's sourced in accordance with certain formal compliance requirements). Crawled data is susceptible to sudden access outages or reformatting of the APIs and/or source data on the supply side,

making the source data suddenly incompatible with the queries made on the demand side. The site could change, could stop working, or could stop existing altogether; there's less control over the data sources and hence more risk.

What is more, the legal terms and conditions clearly stipulate the potential ethical concerns of the data in question (in terms of both privacy and business secrecy). For crawled data, on the other hand, such concerns are not as explicit, but may become manifest in the case of operational usages of this data – say, if clients are putting together new services and/or products based on such data.

A general approach that we recommend is of course to always strive for maximum caution, sensitivity, and care for the ethical concerns involved – specifically those pertaining to privacy. The GDPR framework is intended to ensure a minimum of such concerns being adequately dealt with, in the European context, but there is nothing that restricts researchers from also treating data that pertains to non-EU residents in equally careful ways as we would with EU citizen data.

At the same time, the demand for **replicability of research results** often poses a challenge here, since true replicability often requires access to the original raw data used.

Assessing **data readiness** for social science research therefore must take this into consideration, in parallel with all the other considerations already listed in this report. In addition to ask what one can attain with the data, how accessible it is, and what types of accessibility the data have, how sizeable and manageable they are, how useable and understandable they are, one would have to ask how well the data lend themselves to auxiliary, future uses, and possible replication of results.

This is all very much **governed by licensing**. Twitter, for example, recently updated their means of access for academics, encouraging academic researchers with specific research objectives to apply to their Academic Research product track, providing a richer access to the API than previously – giving, for example, free access to the full history of public conversation via Twitter's full-archive search endpoint, which was previously limited to paid premium or enterprise customers, and also a significantly higher monthly volume cap of 10 million tweets (compared to the limit of 200,000 that the standard API access would have; see Tornes & Trujillo 2021). Nevertheless, as Gold (2020) has noted, the datasets of different online platforms are – unlike many datasets used for secondary data analysis – highly dynamic, ever-changing, and effervescent. "The contents change regularly, not just by the addition of new tweets, but also by deletion and other user-driven changes to the status of available information (tweets, accounts etc.)" (Gold 2020: 5). The supporting documents for the Twitter Developer API make it very clear that "users have control over the public disposition of their data, and

that this should be reflected in its use by others" (Gold 2020: 6). When Twitter users delete or modify the content they choose to share on Twitter, this should be reflected in the datasets harvested by researchers using those tweets for research. This is a policy not entirely unchallenged by researchers, as some discussion is taking place in the research community regarding the balancing of individual users' (perceived) privacy versus the collective needs in society of adequate scientific scrutiny and overview (Sugiura et al. 2017).

These challenges would be particularly pronounced, when it comes to **cross-checking one's archived documents through the available public interfaces of publishers or forum hosts**, long after these documents were originally harvested. If one were to have access to an old dataset and would like to verify it, (e.g., for the purpose of wanting to see if the published results could be replicated, re-using the same data) one would of course want to attempt to identify the provenience of the data used. In such a situation, the availability of the data in the existing dataset could be cross-checked with the current state of the open Web. One could, e.g., peruse CommonCrawl and find these sources or go to a vendor and query for the same data or categories.[7] One practical example would be to identify the individual ID numbers for tweets in a dataset (yes, all tweets come with a unique ID number, which doesn't change even if the user changes his/her username) and use tools to cross-check the status of the related account with the 'live' Twitter feed at the moment of research, since any retained data must be adequately synchronized to the state of the online Twitter data set, as stipulated in Twitter's policies (Twitter Developer Platform 2021), on the grounds of privacy and user consent in particular. This could either be done manually (in the case of a small dataset) or in more automated ways, as Twitter also provides tools for researchers to do so:

> One is an on-demand API for checking the current state of a particular user or tweet (free but rate-limited in terms of requests per time period), the second requires a subscription to the Compliance Firehose, in which Twitter provides real-time updates on content status. […] Researchers intending to accrue Twitter data should ensure that the dataset they intend to collect can be synchronized using one of these methods. This may either incur financial cost to secure the compliance subscription, or may bound the size of the retained dataset to that which can be synchronized regularly through the free API. (Gold 2020: 9)

[7] Note, however, that this shouldn't be confused with the process of identifying users' real identities. In the context of Twitter, such activities go under the name of so-called "off-Twitter matching" – using data from Twitter (and/or elsewhere) to identify or otherwise associate a Twitter user with their identity elsewhere – is restricted, as such matching would requires specific opt-in consent on behalf of the human individuals involved (Gold 2020: 8).

Moreover, **particular infrastructures and competences** may be needed for search, access, and fetching of data. Say, for example, that a computer science or linguistics study has involved a dataset containing 12 million URLs. Theoretically, these are "available" but not readily available or free. Some URLs may be deactivated – indeed, significant parts of the open Web routinely disappear – however, probably most of the links will still be (hypothetically) available. You may then have to go to a provider and repurchase the data. But this is not enough; perhaps resources are needed to systematically cross-check the old dataset with the new one.

Lastly, there is yet another legal obstacle to data provision, namely **different legislative contexts in different countries**. It is obvious that, from Western European and Northern American perspectives, huge regions like the Russian-language market and the highly regulated mainland Chinese are much less accessible, in terms of being able to obtain data. China is, as is already known by most people, home to a range of enormous online platforms: Baidu, WeChat, QZone, etc. Russia, while much smaller as a national market, compared to the Chinese, is known for its similarly highly censored and undemocratic governance of its media and internet markets. Some providers outside of Russia have licensing deals with VKontakte and can thereby access at least partial aspects of this vast social media platform, and Socialgist (worth noting – owned by a Japanese business intelligence company) has managed to obtain licensing deals also with Chinese social media platforms. These types of arrangements are rarities, at the time of writing, and it's unclear to us what the actual quality and scope of the data streams in question are like.

3.2 Different source provenance and linguistic distributions for different providers

Which countries and languages do the respective providers cover more or less well? In order to answer that broad question, we had a closer look at the offerings of a small selection of providers mentioned in this report.

It shall be clear by now, in this report, that several factors need to be considered when using different data providers for online-mediated language data. The original source provenance is something that requires a considerable amount of faith in the original providers, as one would assume that it is in their interest to adequately categorize the content in terms of source provenance. This is a question of **data quality**. As part of the LES project, a useful checklist was compiled, serving as a good heuristic for assessment (Table 2, below).

Moreover, there is also a question of what can be labelled **quality of access**; the technical means, as well as the legal terms and conditions for how the access is made possible in the first place. To begin with, does the analyst get access to the raw data at all or is access only taking place through preformatted dashboards, giving cumulative answers to search queries while not necessarily providing direct access to the source documents in question. We see, that with the transition from API access to "analytics" packages, Facebook's data provision has gone in this direction, *tout court*. This requires a very considerable degree of faith in the providers in question, since the analyst will only ever get second-hand interpretations of the data, never the actual data in question!

Table 2. Checklist for data providers

1. What types of sources do you provide? I.e.:
    A. Open/free web
    B. Press releases
    C. Blogs and social media (e.g.,Twitter, Facebook, Google+, Youtube)
    D. Closed/deep web (e.g., licensed material from Dow Jones, AP, AFP, PRNewsWire, Lexis Nexis, or materials from paywall sites, e.g., Financial Times)
    E. News wires
    F. Broadcast radio and TV as text
    G. Forum, open/closed, pay-to-access
    H. Other
2. What kind of add-on services do you provide, apart from delivering data – e.g., sentiment analysis, named entity recognition, temporal markers, event detection, demographics, geolocation?
3. What languages do you cover?
4. How many sources do you cover?
5. What is the breakdown of those sources across types, e.g., blogs, forums, microblogs, etc.?
6. What is the distribution of sources per language, e.g., 100,000 English sources, 10,000 Swedish sources, etc.?
7. What languages are you planning on introducing next?
8. How do you measure coverage for a given language or socioeconomic/geographic region?
9. What quality metrics do you use internally with respect to the data you deliver?
10. Do you have a Service Level Agreement regarding data quality?
11. Can you add new languages if we ask for it? What will it take from our side?
12. Are (blog/forum) comments included in your sources?
13. How do you identify the language of a source/individual information item?
14. What is the latency from indexing to delivery?
15. How often do you attempt to fetch new information from the sources?
16. How do you handle de-duplication of sources/individual information items?
17. To what extent is the data you provide spam free?
18. How do you avoid spam?
19. Do you maintain an archive of the contents of the sources you have crawled accessible to your customers?
20. Is it possible to obtain development licenses in addition to a full license?
21. What pricing plans do you have, e.g., special deals for start-ups?
22. Who else relies on your data?

In addition to the validity of the sources in question, the analyst might take an interest in h**ow representative they might be for an imagined target population**, as we have already discussed in this report. What would be of most primary interest, to a lot of social scientists, would be key questions regarding the

potential target populations, especially regarding the provision of demographic data (gender, age, possible income cohorts, etc.) for the sources. In terms of editorial sources, this could be an itinerary of the target audiences; for user-generated sources it could be a set of indicative metrics for the original users providing the data. Can anything be said about the geotagging of the sources? Where does such geotagging come from, in that case; who provides that tagging?

Lastly, corpuses are almost always very skewed in terms of the respective publications/source URLs represented, often displaying Zipfian distributions. Here, one could consider several different means of calculating different types of "weight":

You can rank sources by **numbers of documents per source publication**; X amounts of documents from publisher N, and so forth. The drawback of this metric is that it does not take into consideration the vastly varying sizes that different documents can have, nor does it take into consideration the possibility that there might be a high degree of repetition (redundancy) inside the corpus.

You could also rank by **amounts of data – either by numbers of words or in terms of bytes**. This is arguably better than the above measure, however it might also fail to consider the degree of repetitiveness in the corpus; there is often repetition of words or phrases, something that would not be accounted for by the manifest counts of words or bytes. However, by using algorithms for data compression (the most well-known such technique is the compression algorithm used for zip files), also the redundancy can be accounted for.

Alternatively, one could also assign weight to documents by also considering the **estimated popularity of the respective source publications/URLs**. Arguably, this would constitute an important method for making up for the possibility that data models would generally privilege the above parameters, frequency of data. If a corpus would contain very large amounts of text from sources that come from comparatively obscure source publications, in terms of quantitative popularity among the broader public, that would pose a problem: Obscure sources are, by definition, not popular but would still be assigned an outsized weight in the overall models. Hence, a way of assessing relative importance would be to consider a metric for relative popularity or obscurity of sources. In what follows, we will outline a tentative method for doing so.

A small caveat is worth mentioning here: Generally, this challenge of verbosity of prose (Leech 2007: 139–140) versus actual popularity of sources would not be a problem if the data provision was premised on a coincidence in terms of popular sources also being the ones that are the most numerous in the corpuses in question. This is often the case – our Swedish-language dataset confirms this, for example – but should nevertheless not be assumed to always be the *a priori* case.

One simplified way of testing for this popularity dimension is to simply add to our checklist above a question to the providers: **"Could you, for the countries specified, give us your specification of what the top 10 domains are in this country, for this language?"** Meanwhile, the analysts themselves could employ a list of their own, ranking which sources that they see as the top 10. By doing so, the researchers could compare this with the provider's list and, if possible, correlate the two lists. In our section 4 below, we provide tentative such overviews for our 20 selected countries.

3.2.1 Quantitative and qualitative variability of vendor offerings

In a provisional assessment of some of the existing vendors, we found that **corpus outputs from different vendors differ**, not only in that the amounts of available text vary drastically per language, but also in the ways each vendor categorizes its data (in terms of, e.g., tagging, classification of documents, or date range covered) and in the ways in which each vendor makes its data or analytics dashboards accessible: **Some vendors restrict access through giving users access to proprietary dashboards only, while some have more generous APIs that allow for text extraction in bulk.** Needless to say, the latter type of access is risky for data vendors since this makes possible the duplication and redistribution of the data made available. Hence, legally binding contracts are always employed, stipulating what is and isn't allowed to be done with such data; this forms yet another obstacle for researchers to deal with, since, e.g., replicability of results may be severely restricted due to such legal restrictions.

There also seems to be different national biases for different vendors, as regards what languages they tend to focus on. This means that different vendors tend to be able to provide more data from some countries and languages, and less data from others. With such differences in mind, we made a tentative comparison, where language corpora sizes from a small selection of available vendors were put in relation to the actual L1 and L2 populations for different spoken languages in the world, in order to get a coefficient that could provide a rough **"language coverage" rate** that could be comparable across vendor offerings. The metric provided below should be read as highly tentative, as it is based on estimates. For comparison, we have included also CommonCrawl's more recently stated metrics of language distribution (CommonCrawl 2021).

Four competing offerings were put under scrutiny in our overview: the curated selection used for the Lexicon data (consisting of data collected over time from Gavagai), and three other brand-name competitors. Our comparison can of course be criticized from various points of view: First and foremost, it does not compare the different providers on a 1:1 basis, as slightly different parts of each provider's

offering are measured, and these are somewhat different metrics. This is all due to the many restrictions and obstacles to getting totally comparable document counts from each provider.

For the LES Lexicon (Dahlberg et al. 2021b), there was a clear and unambiguous list of numbers of documents in the database per language and per category; it was trivial to compile a pivot table listing the numbers of documents in total for each language. The respective language tagging here was the one originally annotated by the different providers' metadata for the documents constituting the LES Lexicon.

Similarly, CommonCrawl (2021) provides an aggregate listing the percentages of their database covered by each language – measured as the primary language of each html document, as identified by the Compact Language Detector 2 (CLD2) algorithm. This was included as a good benchmark to compare with.

For the Swedish data provider Twingly, a small number of lists were sent to us by one of the company's representatives, tallying the language distributions of the various categories (forum posts, blog posts, etc.) and since this company specializes in blog monitoring and scraping, their most comprehensive linguistic diversity was represented in their compilation of blog posts. We saw their stated count of average numbers of new blog posts per language per day as a useful measure of linguistic diversity, to be put into our overall comparison.

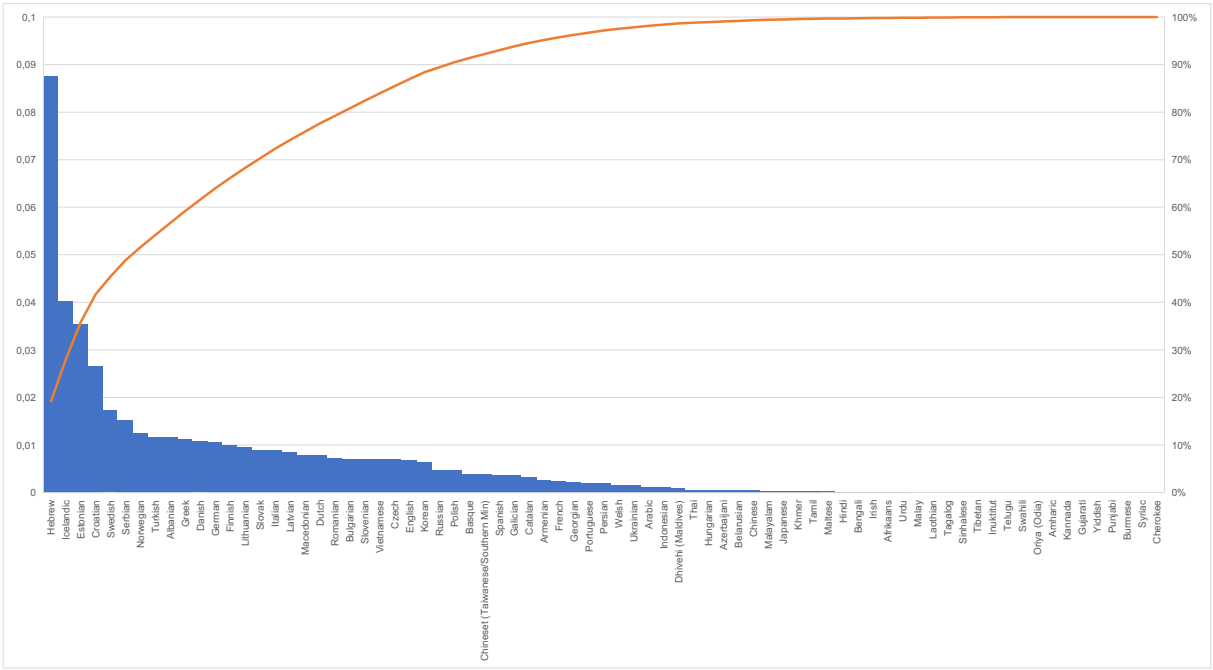
The most cumbersome providers to query for were Talkwalker and Webhose since we did not manage to get full API access to these providers at the time of the experiment. As regards the metrics in question, Webhose and Talkwalker used different synonyms for what we understood as documents: For each language, Webhose stated a particular number of news "sources," which we came to understand as unique documents. Talkwalker used a more confusing term, namely "conversations," which was very hard to estimate in terms of what it actually referred to, especially as there was no information provided by Talkwalker as to how many such "conversations" were kept available in their offering, for each language. So, we found a workaround, where we marked the language chosen and then searched for the most common stopword in that language; a process which we repeated for each subsequent language on our list. This returned a distinct number of "conversations" found in each language containing this stopword – a measure that, despite the shortcomings described here, could be used as a proxy for data size for each language in their database.

As regards the estimated count of L1 and L2 speakers, a compiled list was made using available data sourced from Wikipedia and other comprehensive estimates. This measure, admittedly hard to estimate with total certainty (since a lot of L2 competence is a matter of individual judgment on behalf of the actual speakers).

As long as the same metric was used as a benchmark, we did not consider this as a huge problem – once again, the pragmatic ethos was our primary concern here.

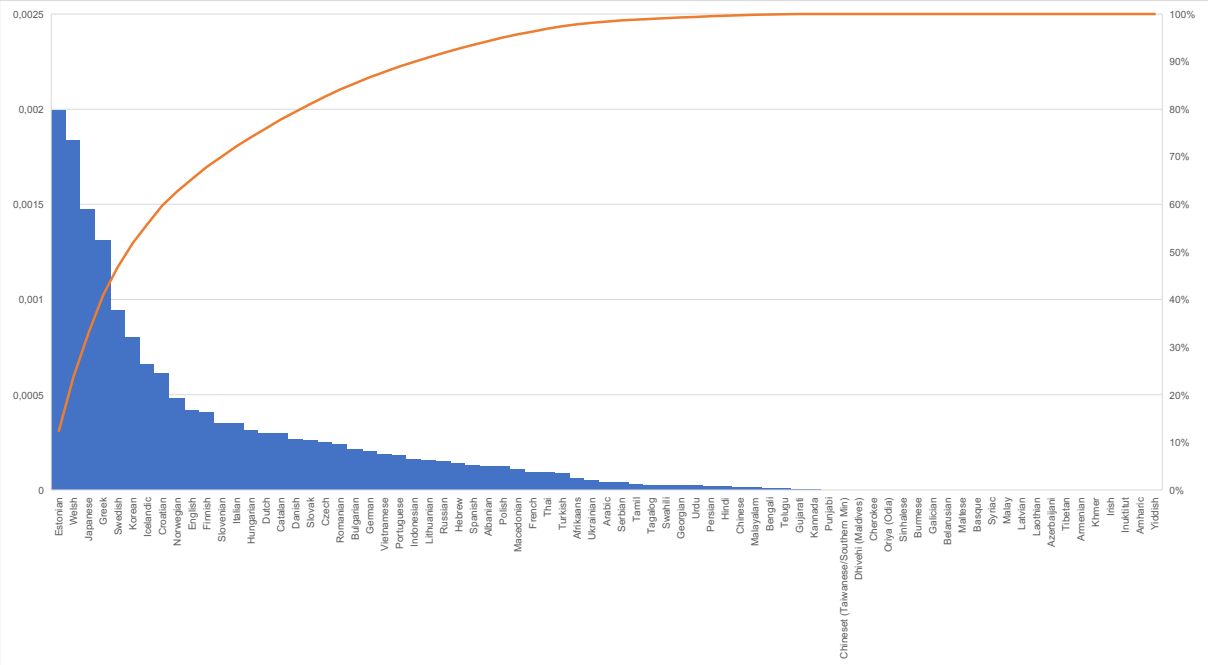
All in all, what is presented here is a **general tendency observed by way of considered estimations of frequencies for each language**. Importantly, this measurement should not be seen as the final word on what is offered by different providers at the time of publication of this report; for some of the providers, the data is a couple of years old and significant change might have been made to their respective business offerings since then. Nevertheless, we find that a lot of general insight can be found by making this comparison. For example, it gives a snapshot of the way language data is not entirely on par with the actual popularity of different languages spoken across the world; in particular, this is a problem that pertains to many African and Asian countries and languages. What is also shown is that considerable differences seem to exist in the respective offerings from the different providers, some of these differences seemingly stemming from specific business decisions and choices of clients and sources. Webhose, as an Israeli-based provider, seems to excel when it comes to offering Hebrew language data, for example. Another notable observation from our compilation is that English is not that overrepresented at all, given how large this language is across the world. Instead, languages like Estonian, Greek, Icelandic, Croatian, and Swedish are very well-represented on the supply side.

Figure 1. Webhose, stated count of news sources, last 30 days (Sept 2018)
Corpus size in relation to estimated language size (L1 and L2 speakers)



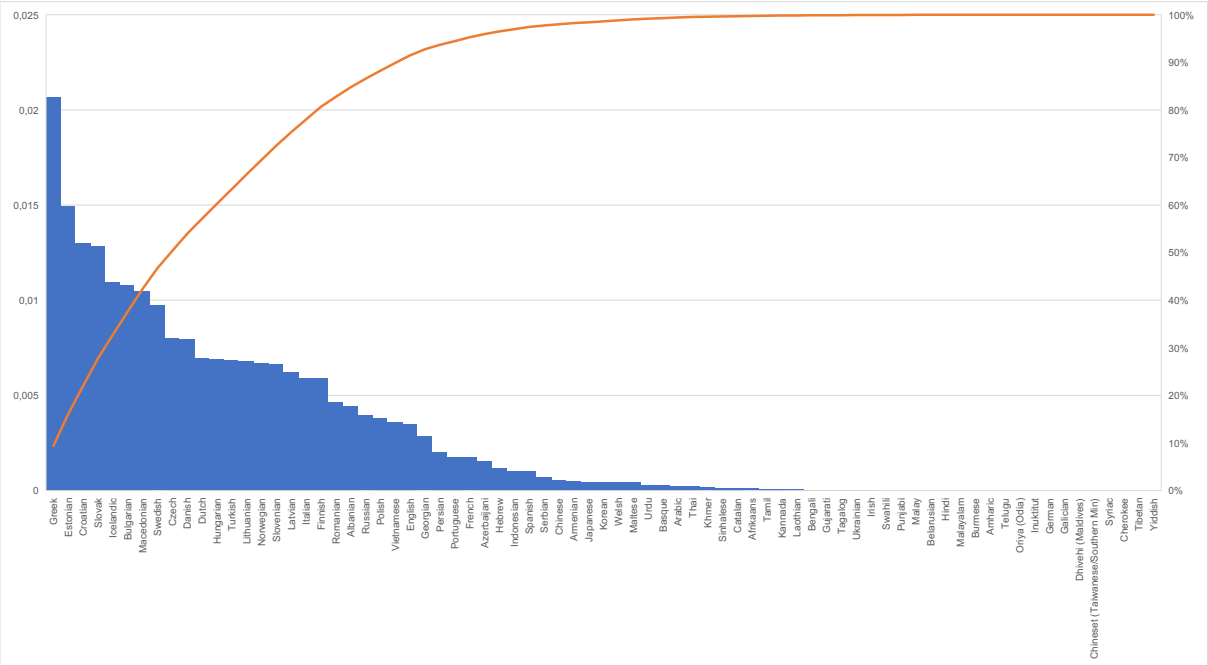
Kurtosis: 27.71
Skewness: 4.71

Figure 2. Twingly, stated average number of blog posts per day for last 3 months (Dec 2018)
Corpus size in relation to estimated language size (L1 and L2 speakers)



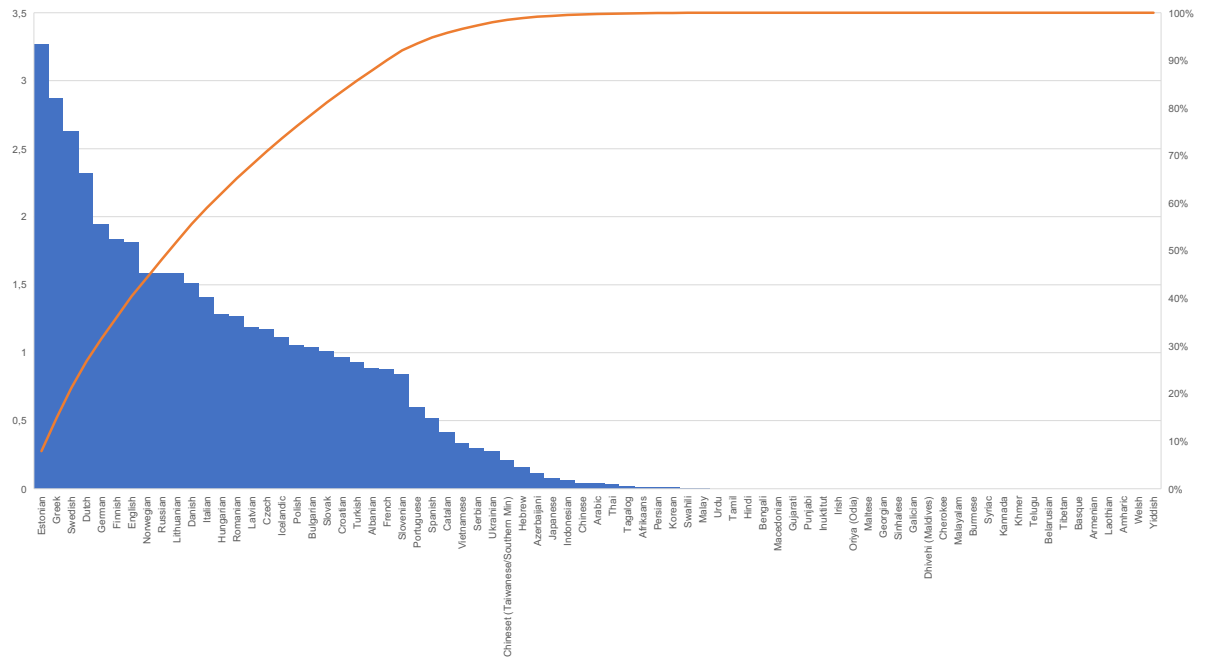
Kurtosis: 9.18
Skewness: 2.96

Figure 3. Talkwalker, estimated number of "conversations" (Dec 2018)
(search query in category NEWS, latest 7 days, queried by most common stopword ["and"] in each language)
Corpus size in relation to estimated language size (L1 and L2 speakers)



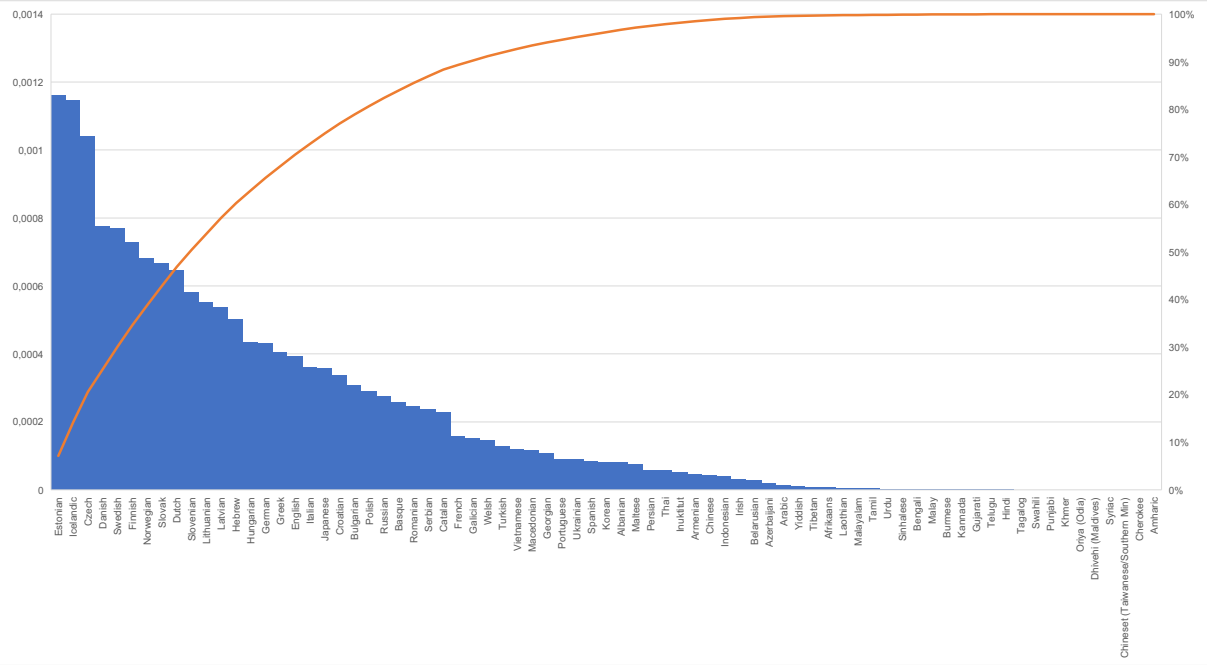
Kurtosis: 3.24
Skewness: 1.79

Figure 4. LES Lexicon, by number count of documents (Sept 2018)
Corpus size in relation to estimated language size (L1 and L2 speakers)



Kurtosis: 1.72
Skewness: 1.50

Figure 5. CommonCrawl, stated language distribution, primary language of html documents (crawl CC-MAIN-2021-21)
Corpus size in relation to estimated language size (L1 and L2 speakers)



Kurtosis: 2.09
Skewness: 1.61

The findings above indicate that the compiled data from several different vendors used for the Lexicon seem to have a less steep variance in language distribution (skewness and kurtosis) compared to some of the other offerings that have more Zipfian distributions in their respective language coverage coefficients (i.e., significantly higher skewness and kurtosis). In comparison, when plotting the currently stated language distribution of CommonCrawl (2021) in relation to the same population numbers of L1 and L2 speakers, the CommonCrawl distribution displays a similarly low kurtosis and skewness.

In other words, while CommonCrawl seems to provide a similarly tolerable distribution of languages, in relation to the actual popularity of these languages in the world, the project members chose to work with the vendor data in question, as it had been compiled over time, with the aim to increase the coverage of smaller languages, in particular. The skewness and kurtosis distributions presented attest to this manual curation helping to improve the variability of the corpus in question. It should be added that the Lexicon's initial, categorization into 'editorial' and 'social' media had been made *ex ante* by the provider, and the data also appears to have been adequately pre-filtered in terms of potential noise and spam.

3.3 Obscure versus mainstream sources

3.3.1 Heuristics for rapid assessment of validity

Like we noted in the introduction (section 1.3), it is of utmost importance to understand how it is unavoidable that analysts make pragmatic considerations and are forced to find reasonable heuristics when dealing with internet-mediated text.

The nature of large repositories of online-mediated text extracted by means of automated processes is that such data is, almost by definition, noisy, fragmented, and heterogenous – but not to such an extent that it would be unusable. By becoming aware of the limitation, pitfalls, and biases of the global market ecosystem of Web scraping, aggregation, bundling, and analysis, one can become better at assessing the face validity of such text.

Below, what we will do is to recommend for **two principal procedures for quality assessment**. It is our recommendation that researchers at least have some way of checking for each of these, when making their rapid assessments of data quality and quality of access. Whether to also filter out content is a later decision, to be made by the individual researchers in their respective projects. One key point we want to make is that content removal ought not to be done *a priori*, without first trying to assess quality. Sometimes, no filtering or content removal might be necessary, as one for example strives to have complete datasets and deem comparability to be reduced if some of these datasets are manually filtered and some are not. This was the mode of reasoning for *not* filtering in Dahlberg et al. (2021b). Our investigations below pertain only to the Swedish-language corpus in that project, since this is our native language that we also have solid domain awareness about – and it would have been unwise to filter only that language and not the others.

Recommendation 1: Checking for validity of sources

When we made a closer look into the "news" category of our Swedish-language data used in Dahlberg et al. (2021b), we tried to assess the endogenous validity of sources contained in the corpus. Below are our working definitions.

We **wanted to include** all sites that can be seen as primarily invested in factual reporting of news and current societal affairs. Here, we include also financial, market-oriented publications, legally oriented publications, specific business sector-oriented news, labour markets and unions. (In our Swedish data, such special-interest oriented sites were marked with an asterisk, so as to differentiate these from more generally oriented news websites.) Also, more ideologically oriented news sites were included, such as religiously oriented and those with

clearly stipulated with ideological orientation (we had as our intention to omit pure propaganda sites, as for example campaign sites for specific political parties or movements, but such sites were rare to come by in the original selection).

What we, however, **did not want to include** are publications seemingly devoted to one or several of the following topics: sport, entertainment/lifestyle, fashion, health, celebrity gossip. These are rarely oriented towards current events and political affairs, beyond specific product launches or entertainment events. We noted that very much of online content on sites with this type of entertainment orientation was of the type "Five tips to get the right style" or "Meet the new Marvel heroes" and could not be deemed to be of societal concern, in the conventional sense.

Specific websites of government authorities, boroughs our councils, or specific universities were all removed. While one could argue that news-like bulletins do get published on sites like these, we are sceptical regarding the partiality of the news producer, since this news provision is not to be seen as editorial as it is rather part of governance. The same proviso was deemed to be at play when it came to websites of individual commercial actors (companies, brands, trade associations), since also these might publish "news" which are in fact marketing dressed up as news bulletins.

Also, singular blogs that were not deemed to have the character of news reporting were removed.

Recommendation 2: Checking for frequency of sources

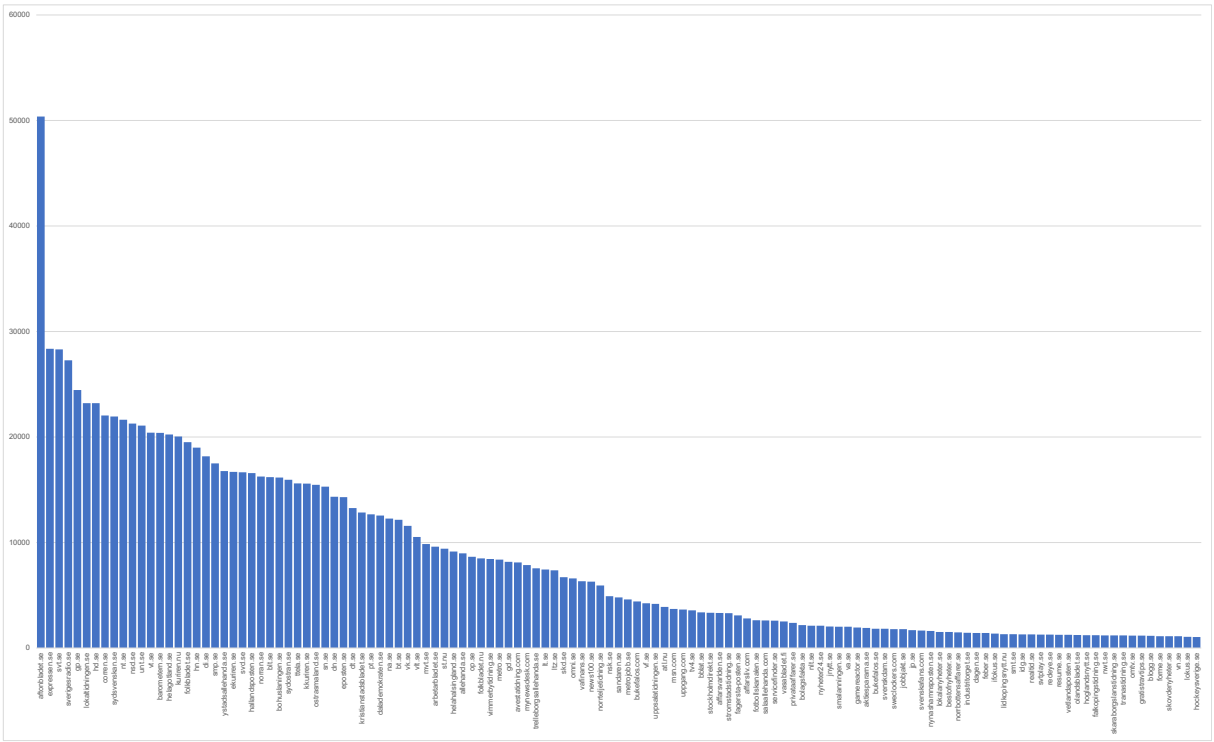
Secondly, since we have noted that the noise-to-signal ratio seemed to increase, the further out in the long tail one moves, a simple decision to improve quality is to **remove outliers**. In our Swedish corpus, we realised that a prudent decision would be to remove all sites with a frequency of less than 100 documents in the corpus. The workload to manually combing through these sites was deemed too high, relative to their frequency. As long as the aggregated number of items in such a long tail is low enough,[8] this is something we generally recommend.

For the Lexicon data (Dahlberg et al. 2021b), we also observed that when tallying the frequency counts of documents in each corpus, the observable distributions

[8] The mode of reasoning, for actors who manage global volumes of data (e.g., Google, Amazon, etc.), seems to be that in those populations, the long tail of obscure content is indeed so long that, aggregated, this content exceeds in volume that of the "head" and the "fat middle" of the distribution curve. It is not for us to say what issues of data quality this quantitative distribution entails, but we want to issue a warning regarding the tendency to lump long tails of possibly invalid sources into metrics like "reach", etc., as some of these global actors seem to do.

are clearly Zipfian in nature. For the Swedish news data, for example, if one includes all 3,000 source URLs in the dataset, out of all these different source URLs, only 357 of them had a frequency of 100 documents or more, and only 127 had a frequency of 1,000 documents or more. There was, in other words, a **long tail of sources,** each source URL only occurring very rarely in the corpus. Among these more obscure sources, the noise was significant; a lot of the source URLs were in fact not news-oriented sources at all; they could be all sorts of websites, representing various commercial, organisational, or individual interests.

Figure 6. LES Lexicon, frequency of documents per source URL (September 2019)
Only showing source URLs with frequency 1,000 or more (n = 127)



On the other hand, among those sources that were much more frequently occurring in the material, the 60 most frequent URLs were all deemed to be valid. Only the 61st URL was deemed to be an invalid source (*news100.se*), and thereafter (for the following 297 URLs) invalid URLs seemed to begin occurring at random. The incidence of URLs deemed invalid among the remaining 297 URLs was as high as 54.2 percent. Including the first 60 (all valid) source URLs again gave a count of 257 source URLs in total; 45.1 percent of these were being deemed invalid. However, when counting the total frequency of documents being deemed invalid, these 45.1 percent of source URLs only represented 93,766 individual documents in total, while the number of valid documents in this manual selection was 1,061,842.

We did this manual selection for all URLs that had more than 100 occurrences in the corpus. In parallel, we also did a A/B split, at a threshold of 2,000 document occurrences, which meant that the top 89 source URLs were kept, and the rest discarded.

We ran a classification algorithm on the manual split, and the result was that while our classifier got an overall accuracy of 0.963 for all data, its accuracy for "good" sources (deemed valid) was 0.986, while its accuracy for "bad" sources (deemed invalid) was 0.782.

In other words, the probability that a document comes from a "good" source, given that the classifier says so, is 0.986, and the corresponding probability for a "bad" source is 0.782. It thus appears that there is a qualitative difference between the texts in the respective documents, and that a classifier may very well distinguish between documents from "good" and "bad" sources, at least to a degree.

Taking the more rudimentary split (weeding out all sources with a prevalence lower than 2,000 documents), the results were similar, but slightly less clear. For this rudimentary split, the classifier got an overall accuracy of 0.947 for all data, its accuracy for "good" sources (deemed valid) was 0.974, while its accuracy for "bad" sources (deemed invalid) was 0.777.

In other words, through using a very quick machine learning-based approach, we discovered that the language in the documents that were weeded out must qualitatively differ from the language in those that were included, to at least some notable degree. While the vast majority of actual documents in the corpus must have been valid and also somewhat representative of the population (at least in terms of reflecting actually popular websites), some parts of the data were likely not ideally to be used as representative of "popular news sites" in the language in question. However, given our discussions of discursive representativity versus linguistic representativity in the chapters above, one might ask whether this makes that much of a difference to the training data.

Nevertheless, for corpuses to be more manageable and, it seems likely, to be more valid and representative, individual URLs could be weeded out, to improve overall data quality. After a certain point, the long tail of obscure sites could be split off, since sources that occur only in the long tail appear to be more likely to be invalid, in terms of quality of content to be deemed worthy of inclusion. It is, however, of utmost importance that cropping and editing of datasets, in this manner, is tested by the researchers involved so that it will not, in fact, diminish rather than improve the quality of data, as was shown in the case of Allen et al. (2021), explored in the section on Facebook curation of data, above (2.4.4). In our own case, the "target population" for the dataset is the external, imagined glut of

online-mediated news prose in general, in the language in question. We wanted to prune the data selection available to us, to weed out obviously irrelevant, invalid specimens. In the case of the Allen et al. (2021) paper, on the other hand, the dataset claimed to cover the target population much more directly; the dataset was thought to be an exhaustive snapshot of the actual population of URL shares on Facebook, but since the very long tail of obscure sources had been removed, this dataset omitted very important information. The devil is in the details.

Summary

In what has been presented above, we have attempted to make a comprehensive overview of the supply side of web-mediated text as, essentially, a data commodity. We have, so far, noted that…

- … not only does the distribution of sources for each language follow Zipfian curve (removing outliers seems to affect the overall quality of the corpus);
- … the corpus sizes for different languages differ vastly (which is, in turn, a proxy for factors like internet penetration in different countries, different national media/publishing legacies, etc.);
- … the validity of the sources found in the categories of "editorial" and, respectively, "social" is sometimes hard to assess (with some document types, e.g., news-oriented blogs, being hard to unambiguously attribute to one and only one of these categories, and some document types, e.g., tech and consumer information, being hard to unambiguously label as politically/economically/socially relevant for social-science research purposes);
- … initial representativeness and degrees of coverage are difficult to assess (i.e., the degrees to which the popular news/media URLs in particular countries/languages are included in the corpus); therefore
- … reliable longitudinal analyses of online media sources would be hard to perform (since this requires consistency of the above measures).

In other words, when it comes to found (i.e., scraped) internet data, conventional statistical methods for assessing validity and representativity are not always suitable, or even available to begin with. This reanimates the already established understanding among social scientists: Contextual knowledge is paramount, as is the human capacity for (abductive) reasoning.

> As Nobel Prize–winning economist Herbert Simon would famously argue, individuals, institutions, and computers have limited information processing, storage, and search powers. Their capacity for "procedural rationality" is "bounded" by computing limits, whether the procedures run on top of human "meatware" or computer hardware. We use rules of thumb and shortcuts to make faster decisions, even if they don't always deliver optimal results. That may be a good thing; chess players consider only the most relevant moves to make due to their enhanced domain expertise. But these "heuristic" programs also may be the result of bias and dogma. If military men who defaulted to bureaucratic "standard operating procedures" had their way, the Cuban Missile Crisis would likely have ended in nuclear holocaust. Many algorithms that deal with complicated or computationally intensive problems use heuristics and shortcuts; the question is whether they are well-chosen. As with all shortcuts, often times they are not. (Elkus 2015)

In earlier work (Bolin & Andersson Schwarz 2015), this challenge is addressed in more academic detail, pertaining to precisely the types of providers that we have already listed: While, facially, 'big data' approaches do not generally build upon established socio-economic variables (since these are often unknown or very badly encoded in the original data), but instead make "blind" inferences from the relational properties of the data – algorithmically recognising which patterns that are most salient (i.e., probable), in terms of co-occurrences (i.e. which data points tend to coincide with which other ones) – "the data mined for pattern recognition privileges relational rather than [conventional] demographic qualities" (p. 1). Nevertheless, "the agency of interpretation at the bottom of market decisions within media companies" is forced to reckon that "heuristics of the algorithm" are always made, anyway and anyhow, since the data always becomes translated back into social categories – because otherwise, the inferences made from 'big data' analyses would have no operational use.

One way of understanding the types of indications offered by ML approaches is that what such approaches always offer are (estimations of) *probability distributions*. By their very nature, such distributions give more credence to general (broad) tendencies since such observable tendencies are often also coming with a higher degree of certainty (probability ranking). Less frequently occurring data points often entail a much lower degree of estimated probability; it is very hard for an algorithm to say things with certainty, about data points that only occur very rarely, while data points that occur with high frequency are generally therefore deemed to indicate much higher probability. By logical inference, it therefore follows that 'big data' approaches in the social sciences are generally oriented towards identifying *broad tendencies* within large and heterogeneous datasets. It is, once this is noted, the task of the human analyst to put such indications in larger societal context.

This is arguably also one of the key reasons as to why "broadcasters and advertisers often remain faithful to the well-worn, scattershot broadcasting heuristic rather than taking the risk of relying on a highly tailored, convoluted process of identifying 'relevant' patterns and then tailoring communication to those fractions of user profiles that emerge" (Bolin & Andersson Schwarz 2015: 7). *Mass appeal* remains functionally relevant, since the degree of uncertainty increases the narrower the targeting becomes!

It is with this knowledge in mind that we will now turn to a few of the practical heuristics for improving the assessments of the data in question. One such heuristic is to use Web traffic data for source URLs as a proxy for popularity. Please note the provisional nature of such an endeavour; as we have noted before, the aim is never to provide an ultimate, impeccably reliable metric of popularity, but,

rather, to find proxies that (faulty as they may be) would provide a common currency by which comparisons can be made, and relative differences therefore be identified.

3.3.2 "Unique users" as a proxy for popularity of Web sources

The idea behind this section is that we can assess (in a somewhat comprehensive and comparable way) the **relative popularity of different text-based online media in different countries**, by noting the relative frequency of Web traffic, as measured by commercial statistics providers like Alexa and SimilarWeb. These actors use proprietary methods to try to measure, as realistically as possible, the actual traffic from unique users that different top-domain URLs see. These actors do so by employing rather sophisticated composites, triangulating different methods.

Services like Alexa and SimilarWeb are equally as commercial as many of the data vendors listed above. Web traffic measurements are a key performance indicator for online businesses of all sorts, central for competitive analysis. The rankings of these two companies are constructed as **composites between how many users are estimated to have visited each site, how many pages have been viewed, and for how long**.

Alexa[9] offers a global rank table, including millions of websites, listed in order of popularity. Through the above composite metrics, average daily unique visitors, and numbers of pageviews for a given site over the past 3 months are estimated. The lower a website's Alexa rank, the more popular the site is. A key component of Alexa's is the aggregated activity of participating users; ordinary web users who install an Alexa toolbar in their private Web browsers. The value for the user is that it displays the Alexa Rank of the visited website. The value for Alexa is that it also sends traffic data to a central server, recording the user's IP address and the URL that the user is visiting. Alexa also measures traffic directly from sites that choose to install a special Alexa script, and the company also has methods of certifying these metrics. By combining metrics from its global data panel and these participating websites, Alexa makes daily calculations of frequency of visits and puts them into relation with the estimated global figures. Users with the Alexa toolbar installed of course constitute an extremely small part of the total web surfing traffic, so this aggregate must be itself treated as a sample, with all the problems of estimating representativity that such an endeavor would entail.

[9] Alexa was founded in 1996 by Brewster Kahle (who also invented the Internet Archive), together with Bruce Gilliat, and was acquired by Amazon in 1999.

SimilarWeb, founded in 2011, works more as a metaservice, synthesizing a lot of different data points, in order to provide an aggregate measure of site popularity. Like Alexa, they have a set of participating websites and apps that share their first-party analytics with SimilarWeb. Moreover, they utilize something they call a "Contributory Network" – a collection of consumer products that provide anonymous device traffic data that is aggregated at the site- and app-level; "Partnerships" – a global network of organizations such as internet operators (ISPs), measurement companies, and demand-side platforms (DSPs) that capture behavioral signals across the internet; and "Public Data Extraction" – an aggregation of online information available to the public, that is algorithmically combined with census data such as country populations, in order to produce comparable estimates (SimilarWeb n.d. a).

The measurements offered by these commercial actors are far from perfect, and critics have, for example, compared them with first party analytics (e.g., Google Analytics) to show how there is a degree of arbitrariness depending on how one weighs the different components of one's index, for example. Different actors might measure time frames differently, they might count devices differently, and have different definitions of what constitutes a "unique visit" or "unique visitor" (more on this below). Google's way of measuring sessions and pageviews seems to calculate a bit differently, and perhaps utilize different conditions as for what should count as a pageview or session. Google has the luxury to be able to include all the sites that have Google Analytics connected to their daily operation.

Another challenge is that SimilarWeb and Alexa generally do not count subdomains and subpages in separate ranking, only top-level domains. The reasons for this are likely to be manifold; one key observation is that measuring subdomains and subpages would open up for a possibly unmanageable complexity, as regards what should count as a separate unit or category, and what should not.

For these reasons, SimilarWeb provides the following statement, in relation to the question about whether direct measurement data might show different results compared to SimilarWeb's aggregated data:

> The majority of businesses use Direct Measurement tools to measure and analyze traffic to their own domains. Although the technology is usually similar from one tool to the next, surprisingly, the data often varies. This is because of different methodologies used to calculate sessions, session time, and other simple and standard metrics. For example, some methodologies deduplicate visits and/or remove bot traffic, and others don't. (SimilarWeb n.d. b)

The company underscores that estimates for sites with small volumes of traffic are hard to create; "for websites with a small number of visits, our estimations

will not be statistically significantly accurate. As a general role, for websites with over 100K monthly visits, will feel very comfortable with our estimations" (Similarweb n.d. b). Importantly, the company adds (and here its statements are in line with what we argue in this report) that there will be variances and discrepancies between different ways of measuring Web traffic, but if an actor offers a methodology that is *consistent* this will create comparability over time and across various sites measured by way of that methodology. For measurements in general, since they are composed by estimates and aggregates, the bigger the sample size, the better the accuracy level tends to be. Data is, in other words, more accurate for the very popular websites than for the very obscure ones.

A key concern, when measuring Web traffic data, is to differentiate between pageviews – which are numerous, since singular users often generate multiple pageviews as they explore different websites – and singular users. So, what would constitute a **unique user** or a **unique visitor**?

First, it is important not to conflate singular web browsers with singular individuals. One person could, for example, interact with the same site or service with several different devices. Second, when users are allocated IP addresses dynamically (for example by dial-up Internet service providers), this may cause metrics to overstate or understate the real number of individual users concerned. Generally speaking, unique numbers of visitors refer to the numbers of distinct individuals requesting pages from a website during a given period, regardless of how often they visit or the frequency and length of browsing sessions.

It is in the interest of standardization agencies that companies use a shared set of global standards of measurement. Notable examples are the Marketing Accountability Standards Board (MASB) and the International Federation of Audit Bureaux of Circulations (IFABC). The latter is a voluntary federation of industry-sponsored organizations, jointly collaborating on formulating standards of measurement, that was instrumental in forming a Global Web Standards Group in 1997, in the service of creating conditions for comparability of metrics across jurisdictions.

Since a visitor can make multiple visits in a specified period, the number of visits may be greater than the number of visitors. When an individual goes to a website on Tuesday, then again on Wednesday, this is recorded as *two* visits from *one* visitor (Farris et al. 2010: 327). A visitor is therefore often referred to as a *unique* visitor or a *unique* user to clearly convey the idea that each visitor is only counted once. However, depending on the time frame of as a unit of measurement, one and the same individual could count as several unique visitors over time, as we shall see below.

"Monthly Unique Visitors is the sum of devices visiting the analyzed domain, within the country and time period analyzed" (SimilarWeb n.d. c). In order not to anthropomorphize this metric, it is perhaps more instructive to label it "monthly unique visits." The inquisitive user discovers that this is in fact also the term that SimilarWeb employs on its dashboards. *Monthly unique visits* are merely counting how many unique device sessions there have been on the page over the course of a month; this does not represent a straight 1:1 count of real people. It is, however, a proxy for reach, which can in turn be understood as a measure of popularity. Once again, the useful dimension is the comparative dimension over time; even if page visit metrics are never exactly corresponding with human beings, they are measured consistently and thus become useful metrics for comparing sites with other sites and comparing the development of traffic for individual sites over time.

Different services define page visits somewhat differently; Alexa defines a "visit" as a single browsing session. "If a visitor views another page on your site within 30 minutes of the last pageview it is counted as the same visit. If a visitor returns to your site after 30 minutes have passed since the last pageview then it is counted as a separate visit" (Alexa n.d.). Other services sometimes employ slightly different definitions, such as "aggregates of pageviews generated by the same user during the same session (i.e., the number of sessions during which that page was viewed one or more times). The time limit for a given session is 24hrs" (Rockcontent n.d.).

By calculating the numbers of visits per unique visitors, one can get a rather good measure of "stickiness" or audience loyalty (Hindman 2018: 36, Jenkins et al. 2013). If a website has a lot of visits per visitor, this can indicate that its users stay on the site and read several different pages on it when they visit. However, this measure can be the result of other factors as well, such as website design and incentives created by publishers to stay on the page or urge people to visit subdomains and/or other pages on the website in question.

In sum, it is instructive that researchers assess the sources that they are exploring in terms of their **manifest reach metrics**, but also that researchers take heed and note potential websites that they discover having significant reach in the country or language community in question, but that is not captured or included in the corpus. Not only should the sources found be assessed in terms of reach (i.e., assumed popularity), but considerations should be made as to whether it is apt to keep the sources found, or to assign weight to them, if the corpus consists of numerous posts, and significant amounts of plaintext. **Should sources be privileged that are wordy and therefore occupying a lot of the data in the corpus – but manifestly obscure?** If a web source is rarely visited by the public and thus enjoying little traffic, as manifested in page visit data such as

this, should this be accounted for in the research methodology? There is no simple answer to this question, and what we have tried to show in the chapters above is that it very much depends on what qualities one aims to capture with one's data. Are we optimizing for discursive (topic) representativity, or for structural intralinguistic representativity?

To get a realistic understanding of the apparent popularity of sources, each researcher should consider websites that he/she is acquainted with as a starting point for comparison with foreign sites. By comparing the observed numbers of foreign websites with those that are closer to home, the researcher can get an understanding of the relative popularity of the sites under scrutiny. Let us thus take our own familiar context of Sweden as an example.

Table 3. Example of listings of traffic data, SimilarWeb

	March 2020		April 2020	
	Total visits	Unique visits	Total visits	Unique visits
aftonbladet.se	101.3 M	7.865 M	90.88 M	7.403 M
expressen.se	86.50 M	8.756 M	77.95 M	8.175 M
dn.se	26.59 M	4.084 M	24.95 M	3.538 M
svd.se	16.97 M	4.036 M	16.00 M	3.618 M
flashback.org	15.18 M	2.385 M	14.09 M	2.252 M

We cannot say that Aftonbladet would inherently have more statistical "weight" from such numbers alone – for one thing, comparing unique page visits for different news/editorial sources is like comparing apples and oranges, since some media have loyal readers who access news and content mainly through dedicated apps, which does not register in metrics like these, and, moreover, each individual page load could be vastly different in terms of the amounts of information per webpage. Dense websites might need fewer page visits to convey much more information than other websites.

What we can say, however, is that metrics like these are mutually comparable as a rough estimate of relative popularity of different websites. We can get some degree of contextual awareness of what sources are popular, for a particular language community. One complicating factor is that the metrics would generally only divulge the total number of site visitors globally, for each site, therefore making it hard to assess whether all or only some of these come from the country that the actual website is hosted in, or where it has its target audience. A lot of media titles have **transnational audiences**, and for particular languages, some media titles might be popularly enjoyed by large scores of expatriates speaking the language in question but not residing in the same country of residence. Most of the media sources that we observe in our overview have mainly domestic (national) audiences, however. **In the case of media that are popularly enjoyed across borders, these metrics do become more complicated, and harder**

to unambiguously assess. We will return to that challenge for some of the individual countries and attendant domestic media titles listed below. In what follows, we will also list some further provisos regarding Web traffic measurement that we noted as we began charting the various online news media in different countries.

4. Countries in context

4.1 Challenges with the country-comparative approach

It is our intention, in this report, to provide **a brief outlook on a selected number of countries**, in order to exemplify how one could approach the many issues of comparability and representativity addressed in this report. We have manually selected 20 countries that we deemed to constitute an interesting cross-section of countries, that each illustrate specific aspects that have been addressed in the above sections, and/or other notable considerations, such as, for example, ongoing political turmoil and/or polarization, and how to deal with that from the perspective of quantitative social sciences. It is with this in mind, that the reader ought to note that we write about some of the countries in significantly more detail than about other countries; the examples of Poland and Hungary, for example, were illuminating in several ways, as these are countries in the absolute vicinity of Western Europe, but that remain partially obscure to many Western observers, probably due to language differences. The political shifts in recent years in these countries meant that some additional contextualization is needed when making an overview of key media titles and current developments in these countries.

Moreover, countries in Africa, Asia, and South America whose media landscapes would not ordinarily be well-known to Eurocentric observers are also covered in some detail, whereas countries like Germany, France, and Italy in our dataset are not covered in nearly as much detail as their popularity of web-based media might demand, in comparison to the much less densely populated online media sphere of, e.g., Nigeria.

Since the Web traffic measurement approach (using SimilarWeb traffic data) seems to generate reasonable results for countries that we are contextually very well acquainted with – websites that we did expect to see in the top league are indeed in the top league, and the order of magnitude of websites map rather well onto other media measurements – we should expect, also when it comes to some of the other countries in our overview, that the data maps pretty well onto the actual media landscapes in each country, providing an accurate measure of popularity for many of the online media sites. Of course, there are nevertheless challenges with this approach, primarily due to several methodological reasons outlined below (4.1.1).

We are very grateful for the very useful contributions from Gothenburg University's Digital Society Project (DSP) and Quality of Government (QOG) dataset, which provide useful data aiming at capturing things like internet penetration and online service provision, internet censorship, online media fractionalization. From the perspective of comparative social science, data has been compiled also on less media-specific parameters, e.g., regime types, degrees of political polarization and corruption, and the state of freedom of expression in the countries in

question. Valerie Caras has helped compiling some key performance indicators and general observations of the various countries. Moreover, Jesse Salazar, affiliated to Södertörn University, has contributed with some of the media overviews, data compilation, and analysis.

4.1.1 Three important provisos for Web traffic measurement

In our work, when mapping individual Web sites and plotting them according to their estimated traffic data in SimilarWeb data, we found several potential error sources. There are at least three provisos that must be introduced before going through the individual countries in our analysis:

1. **Traffic data not capturing in-app traffic**
   It is now commonplace that publishers and media companies provide specific apps for their news services and other services; it's inherently hard for measurement companies to include such metrics in their overall Web traffic metrics. Even if in-app news reading and browsing was possible to include, how should such traffic be counted alongside the already rather elaborate estimates that, e.g., Alexa and SimilarWeb offer for the open Web?
2. **Traffic data not capturing legacy content published on social media platforms**
   In numerous countries across the world (e.g., Arabic-speaking countries), it is common for news publishers to utilize platforms like Facebook and WhatsApp as publishing platforms, so that news bulletins are published natively on these platforms in parallel with the native websites of the publishers in question. Such parallel publishing (outsourcing the distribution, one might say) generates another dilemma of measurement: It becomes hard to know how sizable the audiences are that peruse media content on third-party platforms, without ever even visiting the native editorial website.
3. **The use of VPNs and proxies among media users**
   In many countries across the world – especially such countries where government repression is high and media audiences largely young and technologically savvy (e.g., Egypt), it is highly likely that a lot of Web surfing takes place through cunning setups on the demand side, namely VPN browsers and DNS proxies, innovations that act to obfuscate and hide Web surfing patterns from external surveillance. Such uses are likely to also complicate the measurement of Web traffic, conducted by companies like SimilarWeb.

Another proviso is how traffic, for certain sites, comes from an assortment of countries outside the nominal country of residence for the site in question. For

example, with news aggregators like the Swedish Newsner.com, a lot of its traffic comes from users in other countries than Sweden, and a lot of its incoming traffic comes from social platforms.

Transnational traffic adds to the complexity of interpreting traffic data. What we are comparing for all our sites listed in this report are global traffic volumes for each URL, since comparability would otherwise be compromised; it's not sure whether just limiting the traffic data to one country divulges enough of the popular 'weight' of the media title in question. Consider the BBC's online news; only 10 % of its total traffic comes from UK sources, according to SimilarWeb, while 38 % of The Guardian's traffic comes from the UK. Denominations like these might appear indisputable, but for South China Morning Post, 25 % of traffic is said to come from the US, 17 % from Hong Kong, 7 % from Singapore – and virtually none, or very little, for China. This is an odd pattern. As it turns out, a lot of traffic that appears to come from the US is due to VPN traffic in the South China Sea region (see the section on Hong Kong, p. 135). If we list the titles based on the share of traffic established to come only from Hong Kong, it appears as if titles like China Times, Epoch Times, and Apple Daily are very marginal ones. This is unfortunate since these are very considerably popular titles in the region. Consequently, when tallying traffic data for titles popular in city-states like Hong Kong, the observable traffic for these titles is very high, much higher than it would have been if it was only traffic from Hong Kong being counted. At the same time, if total traffic is not accounted for, weird aberrations might appear that can be somewhat arbitrary.

Mobile traffic adds to the complexity of interpreting traffic data. In countries like Indonesia and Brazil, mobile phones are often the only way for citizens to access the internet. In large parts of the world, outside Europe and Northern America, mobile phones are in fact the primary means of internet access – sometimes the only means of access. In a 2019 Pew Research survey (Silver et al. 2019) of 11 different emerging markets, a median of 53 % across those surveyed said they have access to a smartphone capable of accessing the internet and running apps.

Regarding the increasing importance of mobile devices, another really important factor to take into consideration is that aggregated figures for page visits from both mobile and desktop devices tends to overcount audiences, since the numbers often contain duplicate entries for one and the same user logging in through both mobile and desktop. For this reason, a premium feature that measurement companies have begun employing is so-called deduplication where this error is approached and estimated.

All of these provisos ought to be duly noted and are arguably extra important when it comes to a subset of countries under scrutiny; countries in the global

South where a lot of Web surfing happens through mobile devices – e.g., Egypt, Nigeria, Kenya, Brazil, Indonesia, Malaysia – but also in richer countries where populations are notably repressed and government censorship very high – e.g., Russia, Hong Kong.

4.2 20 selected countries in comparison

Our selected countries are (in alphabetic order) Brazil, Egypt, Estonia, France, Germany, Great Britain, Hong Kong, Hungary, Indonesia, Italy, Kenya, Malaysia, Mexico, Nigeria, Poland, Russia, South Africa, Spain, Sweden, and USA. Our primary aim with this chapter is to be able to answer basic descriptive questions:

- **Which editorial news media sources are the most used ones in this country and/or language?**
- **Which social or user-generated media are used in each country and/or language, that seem to be available for scraping?**

Potentially, overviews like these would be helpful when researchers cross-check with providers (such as the ones listed in the chapter above) to see how well each provider covers the popular media titles found for each country.

Our secondary aim is to provide a comprehensive case study of some of the political and historical issues that are of concern to both social scientists, policymakers, and language and media researchers in general: **What about countries that are undergoing radical shifts in their political and media landscapes?** Here, we include, e.g., countries like Hungary, Poland, Hong Kong, Russia, and, to some extent, also the US. What can be gleaned from a cursory quantitative overview like the one we provide here? Will certain patterns or insights emerge – such as observations of what the actual popularity of certain media titles are, judging from Web traffic data?

A third aim is **to be able to validate the coverage provided by online text data providers** like the ones outlined further above: What large media titles might be missing in the data coverage provided for specific countries or languages? Since the inclusion of specific media titles in a provider's offering is often a matter of licensing and/or agreement, it is sometimes the case that large media titles are missing in the data provider output. Having "blueprints" of the actual media popularity in specific countries can help researchers assess the severity of such omissions.

These 20 countries are selected partly for pedagogical purposes – we want to show how various caveats arise once one maps some general outlines of the political situation and media landscape in a country – and partly out of analytical purposes – we are genuinely curious about the situation and landscape in many of these countries, since some of these countries have seen rather dramatic historical events in recent years, and exhibit tendencies that can be found also in other jurisdictions. For our general orientation regarding these 20 countries, various research databases and transnationally comparisons have been useful. We want to extend our gratitude to the Quality of Government database, World Values

Survey, Reuters Institute, Reporters Without Borders, and Transparency International's Corruption Perceptions Index.

4.2.1 Macro indicators versus country-specific observations

Thanks to colleagues at University of Gothenburg, who have compiled various documentation of nation-level indicators, and thanks to our own measurement efforts using SimilarWeb, we can summarize a range of observational insights. Their period of interest has been the last decade (2011–2021) while our SimilarWeb data only covers the here and now (late 2021).

First and foremost, it is important to note that the indicators we have for macro level developments per country (e.g., comparative metrics regarding media penetration, quality of government, freedom of speech, etc.) are compiled as part of ongoing projects outside the scope of our own project, and they are only intended as very general markers of broader tendencies.

The link between macro indicators of this sort and the empirical reality "on the ground" is of course a sometimes tenuous one, as it is not always clear whether macro indicators capture the complexity of social reality and measure the relevant factors.

How to interpret the tumultuous contemporary changes, in recent years, brought about by stark rightwing populist rhetoric, in some of the major economies across the world? In Brazil, for example, voting is mandatory, and the electoral system is not as convoluted as that in the US (where the tradition of geographical districts and an electoral college affects the demographic representativity of votes). Electronic voting in Brazil is largely standardized and with high degrees of accuracy. Hence, campaigning is fundamental to the Brazilian political system, precipitating significant degrees of mediated persuasion campaigns, especially among the poor. Bolsonaro wants to remove the current system to be able to manipulate results; like Trump, he argues for drastically reshaping the electoral system and argues that there will be cheating and that he will not accept results if he does not win the next election.

Some tentative examples from our country-specific overview:

- Asking a contemporary Nigerian respondent about "social media" ought to include the caveat that the country is seeing significant repression, the Buhari government practically banning social protest in the streets and trying to suppress and even censor social media. This means that if a respondent is asked whether "social media" are, in his/her opinion, expedient

to freedom of speech this very much becomes a matter of whether one means "social media" as a general force in the world, or the situated social media spaces and places for Nigerians in particular.

- Asking a contemporary Brazilian respondent about "media influence," it is worth noting that Globo has historically had a huge influence on who becomes elected, since presidential TV debates are very important. Nevertheless, Bolsonaro's campaign in 2018 was run completely without his participation in TV debates but instead focused on social media –Facebook-owned social messaging platform WhatsApp in particular. Historically, Globo was instrumental to the military dictatorship but also to the new democracy after 1985. Arguably, the media network has had an opportunistic approach, in that it appears to have been supporting those in power, quite regardless of ideology/regime. Except now, since Bolsonaro is hostile to the entire established media system and has created his own channels together with Grupo Silvio Santos and tv broadcaster Rede Record, political actors with considerable support from clergies and churches.

- Asking a Hungarian about "freedom of speech" ought to include the caveat that while Hungary is seeing one of the most widely spanning spectrum of positions in its national political discourses, many of the popular mass media are, in practice, tied to the ruling Fidesz government and are likely to be severely restricted in their freedom of speech. The proponents of said government are of course arguing the opposite, often alleging that the voices that criticize government policy are controlled by some kind of (transnationally spread) hegemony. While, nominally, this means that the data indicates that there is high polarization in Hungary, it is not entirely clear whether this means that there is diversity of opinion or not. It could either be that the fragmentation and polarization of discourse is indeed an indication of diversity of opinion, but it could equally be said that the discourse is forced into a very one-dimensional axis of polarity, basically that most discourse boils down to the above one-dimensional conflict between supporters of Fidesz and its alleged counterpart.

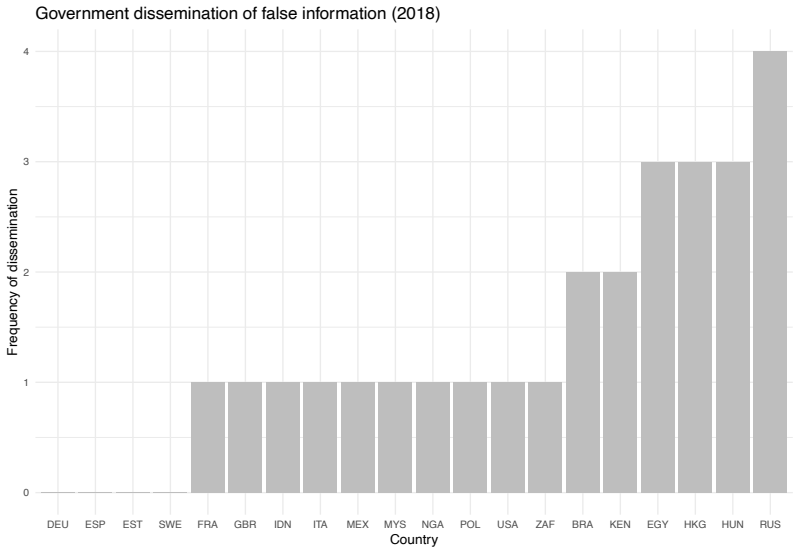
It has been emphasized, by e.g., Strömbäck (2021) how comparative social-science surveys tend to use very broad categories, not only when assessing general societal concepts like "democracy" but also when assessing "media" or "news media." Most research, Strömbäck points out, has focused on news media at the general or institutionalized level, asking people about their trust or confidence in unspecified media such as "the press" or "the media", or asking them about different media types such as "newspapers" and "television". Daniller et al. (2017) have shown how people express much lower trust when asked about unspecified media compared to when asked about specified media.

In the European Values Study, for example, people are asked "how much confidence" they have in a set of institutions and organizations, among them "the press" and "social media" (Strömbäck 2021). While conceptual breadth, like this, is necessary for comparative analysis to be made, our report at hand will show how – hypothetically – different national populations might mean rather different things, since national media landscapes sometimes has very specific characteristics, especially if we consider vastly different countries.

In what follows below, the various graphs compiled by Sofia Axelsson and Valeria Caras are presented and expounded upon, significant parts of the analysis authored by Jesse Salazar.

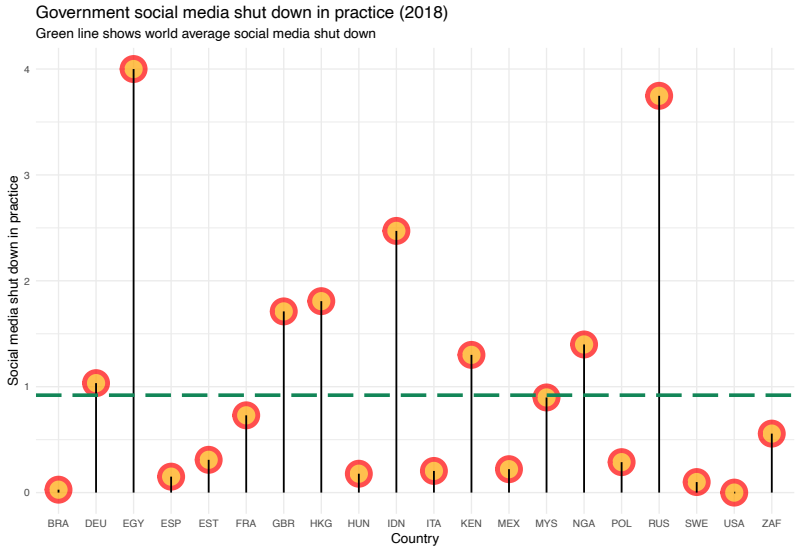
4.2.2 Comparative data on different countries: graphs and analysis

1. Government dissemination of false information domestically



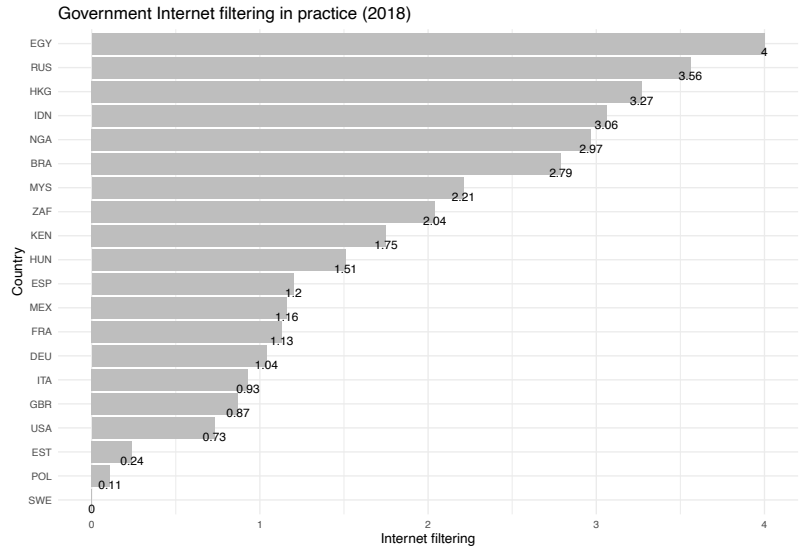
Note: Higher scores indicate a more frequent spread of false information by the national government.

2. Government social media shut down in practice



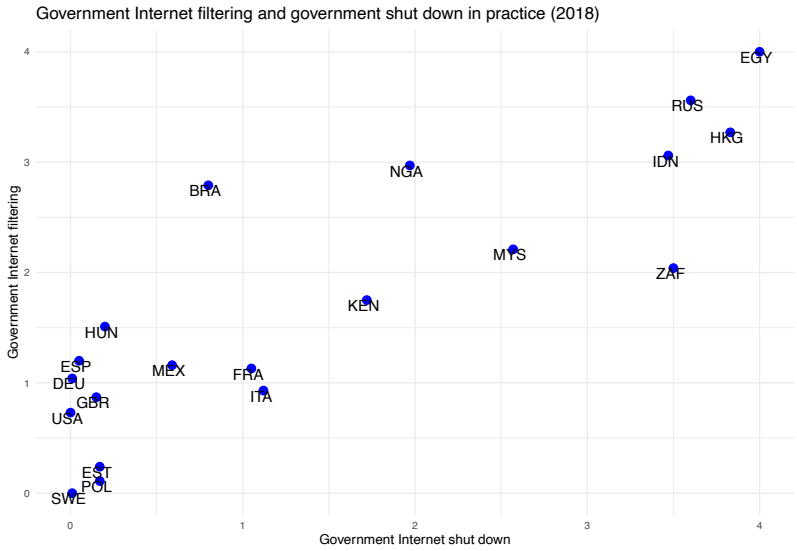
Note: Higher scores indicate a more frequent shut down of social media in practice by national government.

3. Government Internet filtering in practice



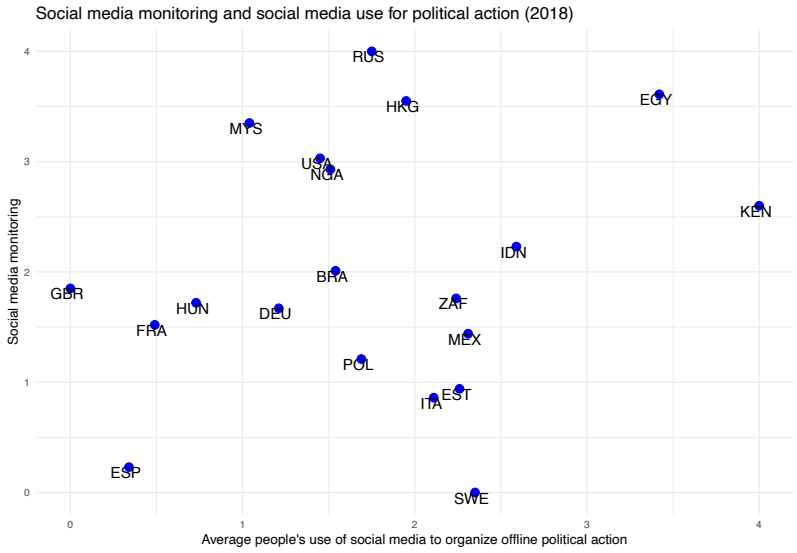
Note: Higher scores indicate a more frequent Internet filtering in practice by national government.

4. Government Internet filtering and Internet shut down in practice



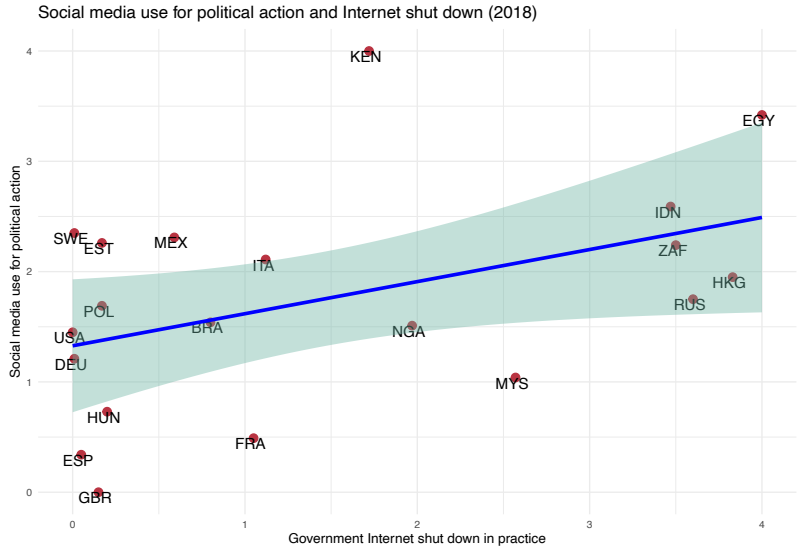
Note: Higher scores of both measures indicate more frequent filtering and shut down of the Internet in practice by national government.

5. Government social media monitoring and use of social media to organize offline political action



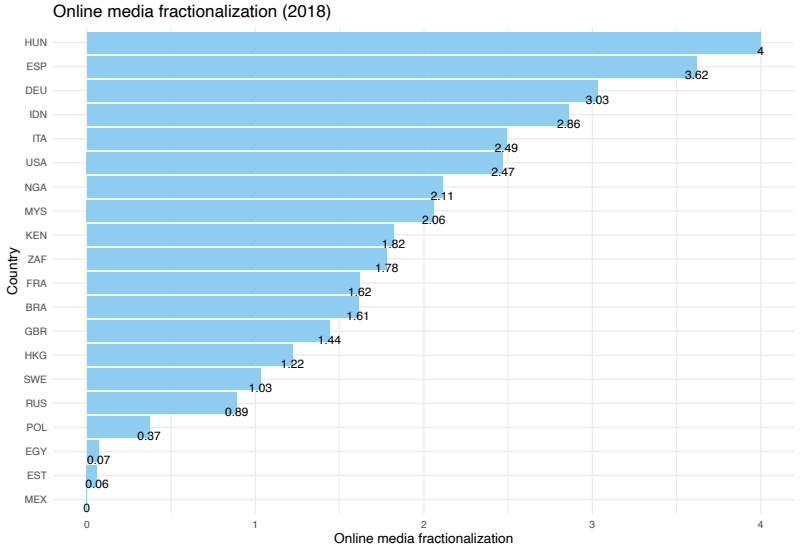
Note: Higher scores of both measures indicate more frequent monitoring of social media by national government, and more frequent use of social media by people to organize offline political action of any kind.

6. Use of social media to organize offline political action and government Internet shut down



Note: Higher scores of both measures indicate more frequent use of social media by people to organize offline political action of any kind, and more frequent shut down of the Internet in practice by national government.

7. Online media fractionalization



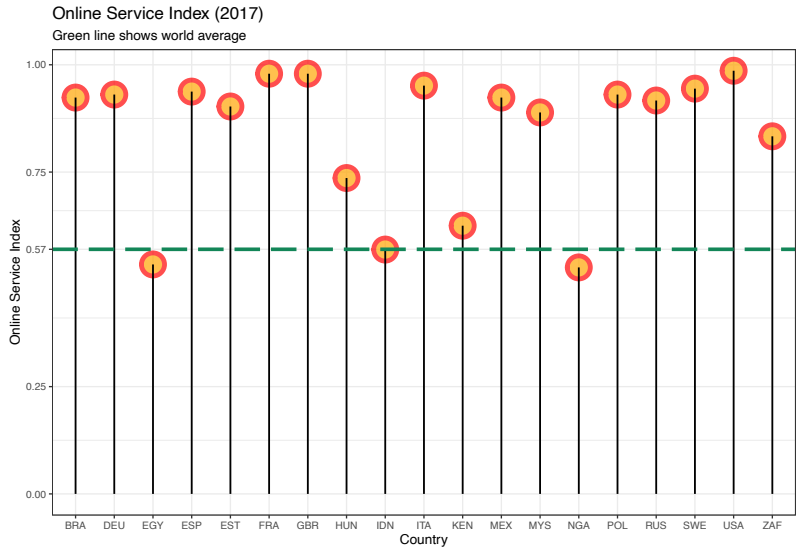
Note: Lower scores indicate more similar presentations of major political news by major domestic news outlets.

8. Polarization of society

Note: Higher scores indicate greater polarization, i.e. differences in opinions on key political issues in society.

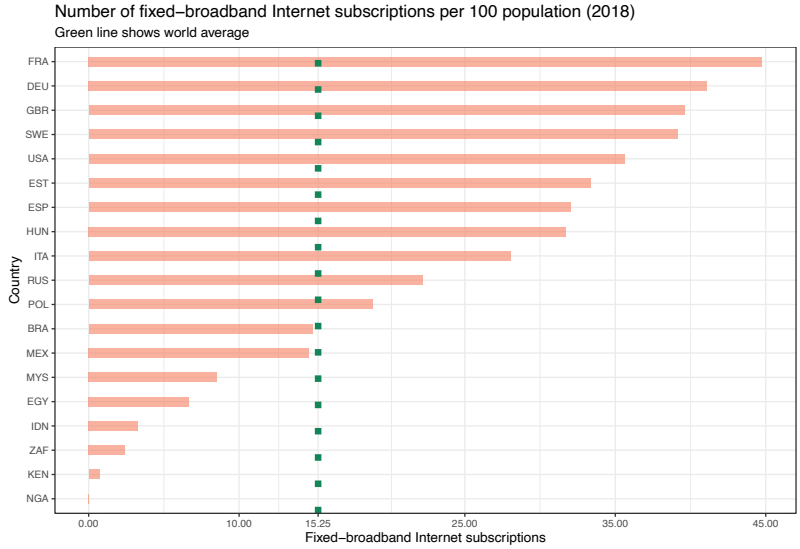
Graphs 1–8 come from the **Digital Society Project Dataset**; Mechkova, Valeriya, Daniel Pemstein, Brigitte Seim, and Steven Wilson. 2020. DSP [Country-Year] Dataset v2. Digital Society Project (DSP). Graphs 9–16 come from the **Quality of Government Standard Dataset**; Teorell, Jan, Aksel Sundström, Sören Holmberg, Bo Rothstein, Natalia Alvarado Pachon & Cem Mert Dalli. 2021. The Quality of Government Standard Dataset, version Jan21. University of Gothenburg: The Quality of Government Institute. Graphs 17–21 are based upon combinations of these two datasets.

9. Online Service Index



Original source: UN Department of Economic and Social Affairs. Note: The index measures each country's national website in the native language, including the national portal, e-services portal, and e-participation portal, and websites of related ministries of education, environment, finance, health, labor and social services if applicable. Lower scores indicate lower online service.

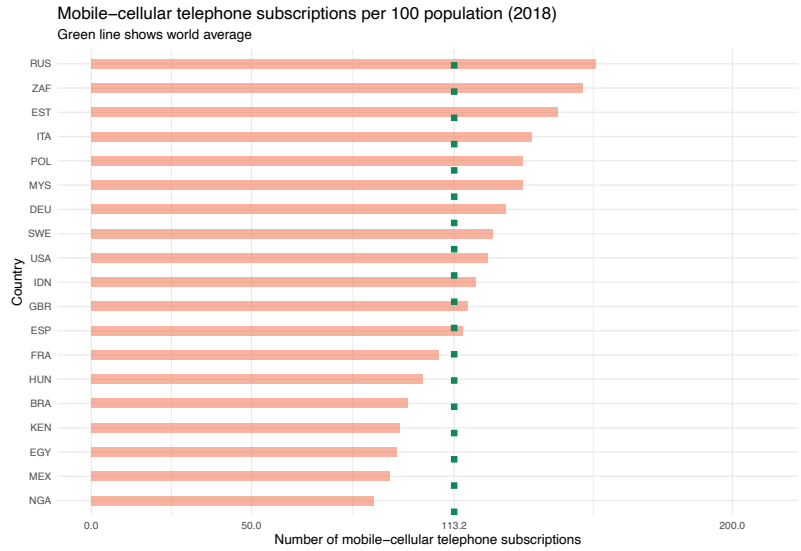
10. Fixed-broadband Internet subscriptions per 100 population



Original source: International Telecommunications Union (ITU). Note: Lower scores indicate fewer fixed-broadband Internet subscriptions.

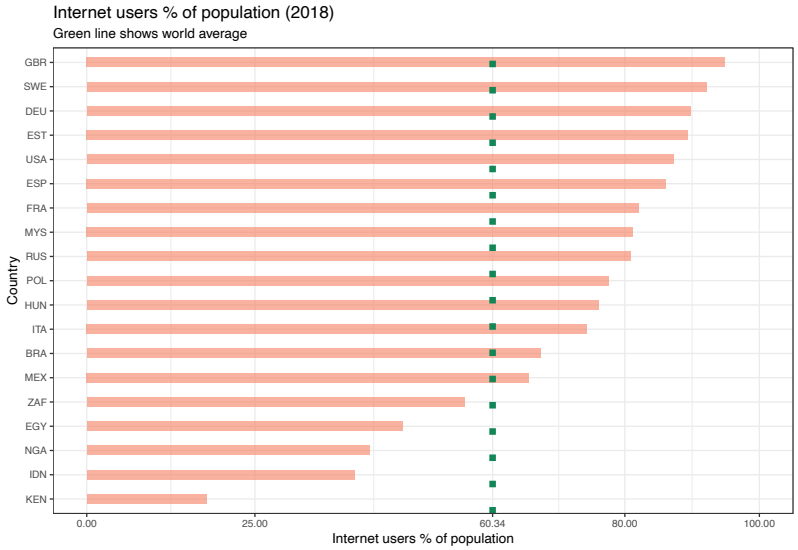
It becomes obvious, when visualizing comparative country-level indicators, that discrepancies between countries that we identify further below in this report (e.g., regarding traffic rates of online text-based news media per capita) are differently distributed than other indicators. When it comes to mobile telephony subscriptions per capita (Graph 11), for example, this is more evenly distributed in the world than the prevalence of internet usage per capita (Graph 12) and our estimates of online news traffic per capita (Graph 32, further below).

11. Mobile-cellular telephone subscriptions per 100 population



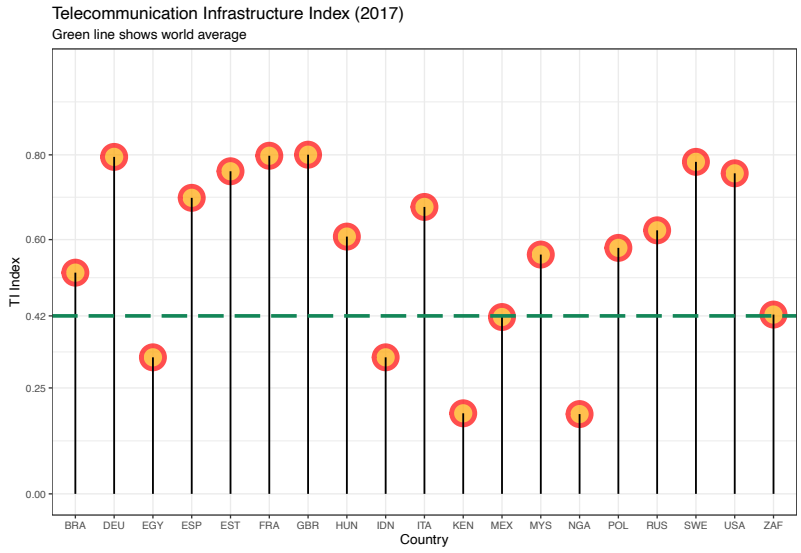
Original source: International Telecommunications Union (ITU). Note: Lower scores indicate fewer mobile-cellular telephone subscriptions.

12. Internet users % of population



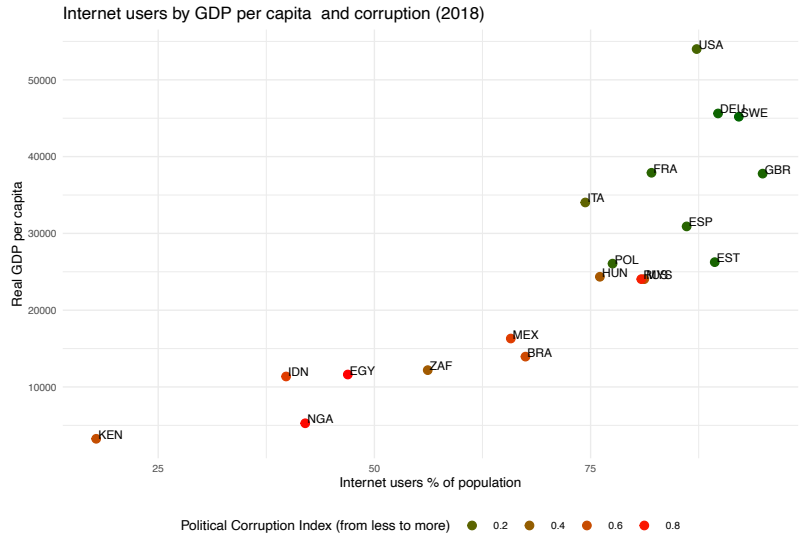
Original source: International Telecommunications Union (ITU). Note: Lower scores indicate lower percentage of Internet users.

13. Telecommunication Infrastructure Index



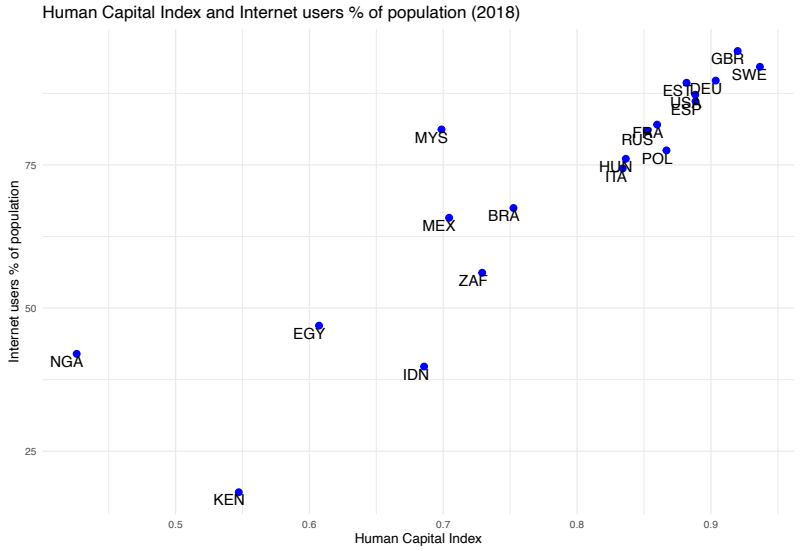
Original source: UN Department of Economic and Social Affairs but with the International Telecommunications Union (ITU) as primary data source. Note: The index measures each country's number of Internet users, number of main fixed telephone lines, number of mobile-cellular telephone subscriptions, number of wireless broadband subaceriptions, and number of fixed-broadband subscriptions per 100 population. Lower scores indicate lower telecommunications infrastructure.

14. Internet users by GDP per capita and Political Corruption Index



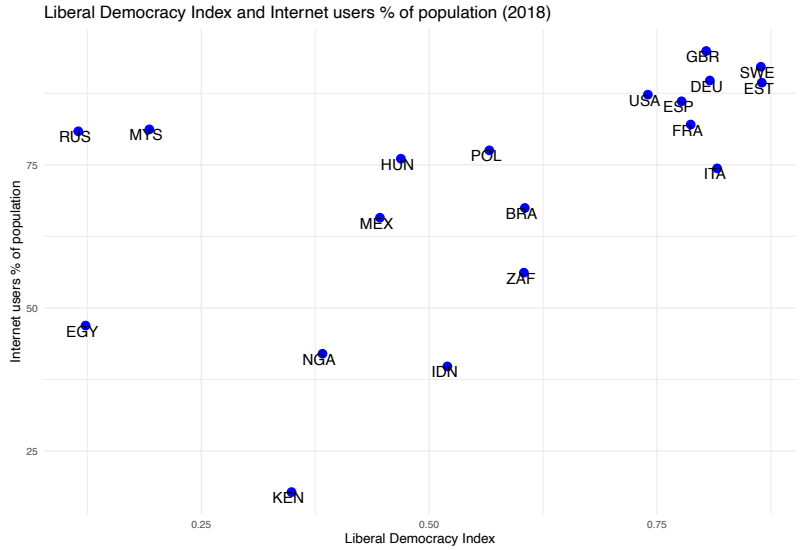
Original source: International Telecommunications Union (ITU) (Internet users % of population); Maddison Project Database (GDP per capita); Varities of Democracy Project (V-Dem) (Political Corruption Index). Note: The graph shows a positive relationship between economic development and percentage of Internet users, and a negative relationship between economic development and political corruption.

15. Human Capital Index and Internet users % of population



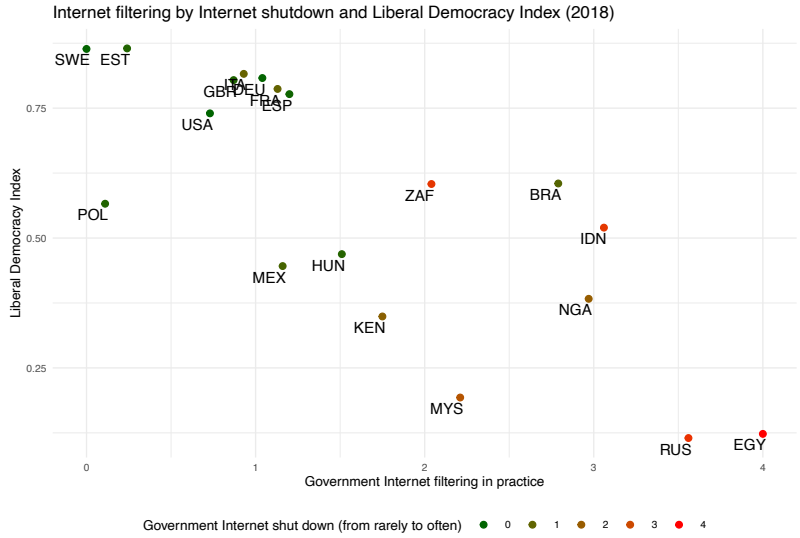
Original source: International Telecommunications Union (ITU) (Internet users % of population); UN Department of economic and Social Affairs (Human Capital Index). Note: The Human Capital Index (HCI) consists of four components: adult literacy rate; the combined primary, secondary and tertiary gross enrolment ratio; expected years of schooling and; average years of schooling. The graph shows a positive relationship between the HCI and percentage of Internet users.

16. Liberal Democracy Index and Internet users % of population



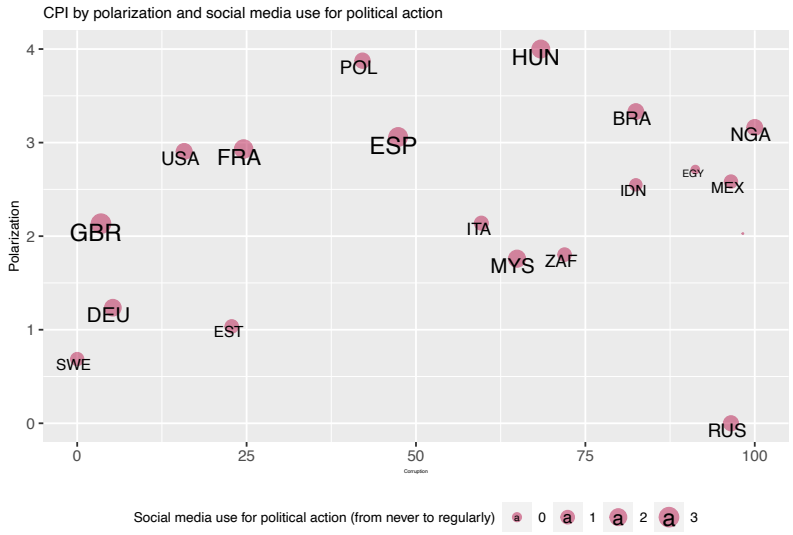
Original source: International Telecommunications Union (ITU) (Internet users % of population); Varieties of Democracy Project (V-Dem) (Liberal Democracy Index). Note: The Liberal Democracy Index measures the extent to which a country has achieved constitutionally protected civil liberties, strong rule of law, an independent judiciary and effective checks and balances that limit the excercice of executive power, and the level of electoral democracy. The graphs shows a general positiove relationship between the Liberal Democracy Index and percentage of Internet users, albeit with a few outliers.

17. Internet filtering by Internet shutdown and Liberal Democracy Index



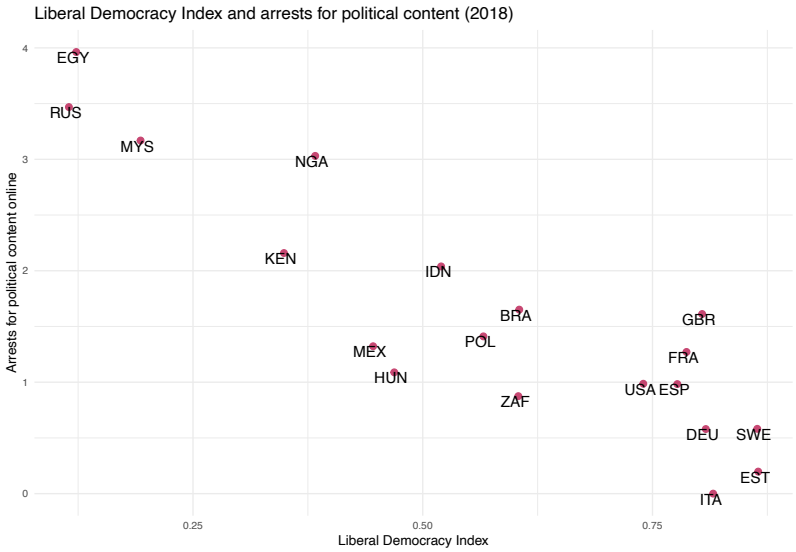
Original source: Digital Society Project (government Internet filtering in practice; government Internet shut down in practice); Varieties of Democracy Project (V-Dem) (Liberal Democracy Index). Note: The Liberal Democracy Index measures the extent to which a country has achieved constitutionally protected civil liberties, strong rule of law, an independent judiciary and effective checks and balances that limit the excercice of executive power, and the level of electoral democracy.

18. Corruption Perceptions Index by polarizarion and social media use for political action



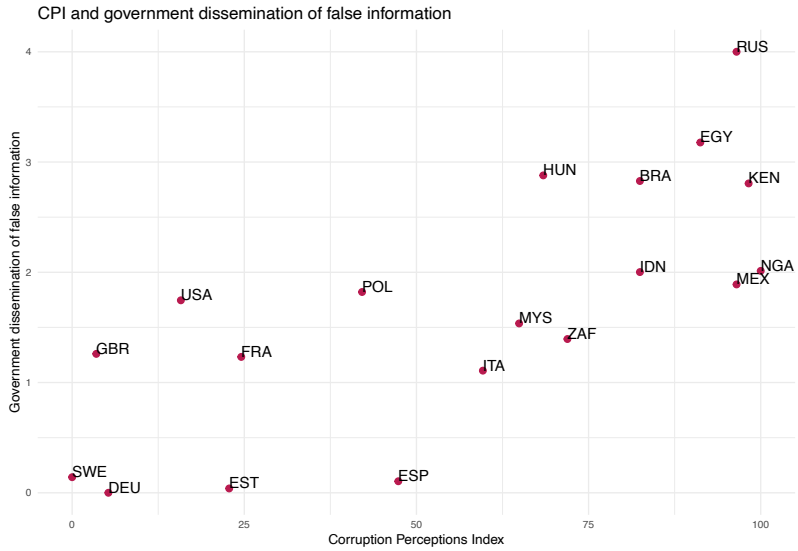
Original source: Transparency International (Corruption Perceptions Index (CPI), measure from 2017); Digital Society Project (polarization of society; use of social media to organize offline political action of any kind, measures from 2018). Note: The graph shows that countries with lower scores on the CPI are in general more polarized on key political issues in society. In more polarized societies, the use of social media for political action is in general higher. Russia appears to be a significant outlier.

19. Liberal Democracy Index and arrests for political content online



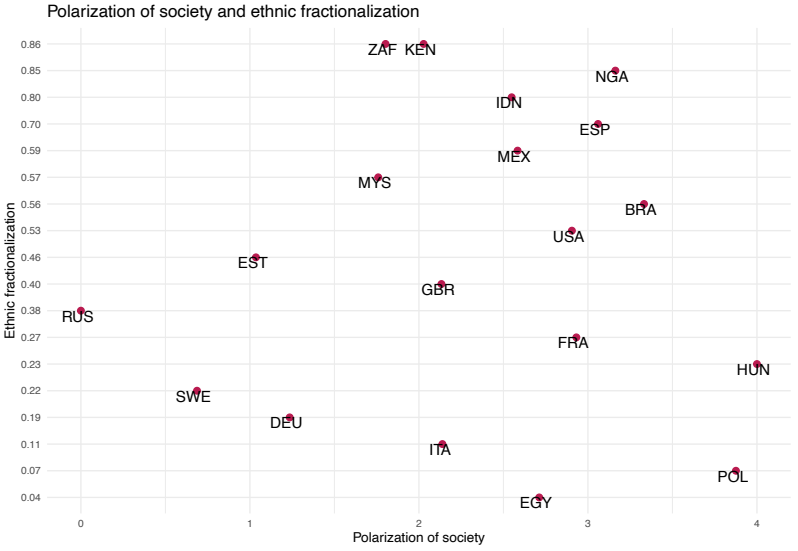
Original source: Digital Society Project (likelihood of arrests for political content online that would run counter to the government); Varieties of Democracy Project (V-Dem) (Liberal Democracy Index). Note: The graph shows a negative relationship between the Liberal Democracy Index and the likelihood of arrests for political content online that runs counter to the government.

20. Corruption Perceptions Index and government dissemination of false information domestically



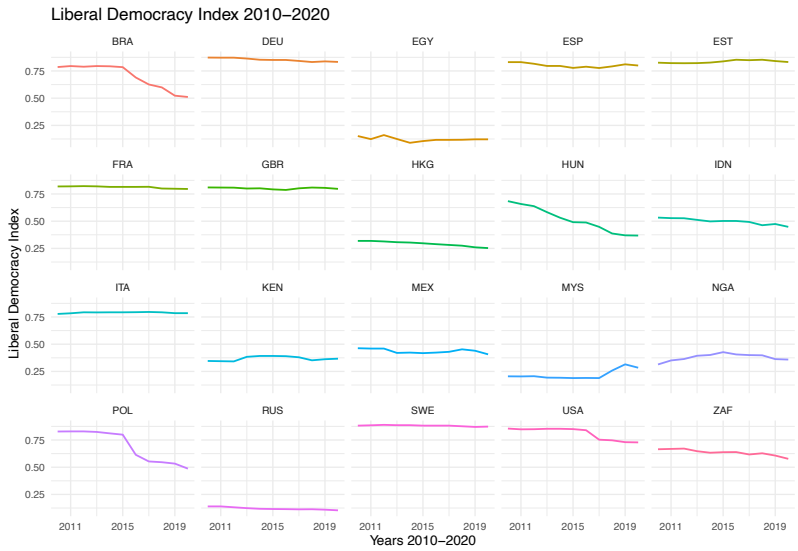
Original source: Transparency International (Corruption Perceptions Index (CPI), measure from 2017); Digital Society Project (government dissemination of false information domestically, measure from 2018). Note: The graph shows a positive relationship between the CPI and government dissemination of false information domestically.

21. Polarization of society and ethnic fractionalization



Original source: Historical Index of Ethnic Fractionalization Dataset (HIEF), measure from 2013); Digital Society Project (polarization of society, measure from 2018). Note: Ethnic fractionalizarion measures diversity as a steadily increasing function of the number of groups in a country, including linguistic and ethnic diversity.

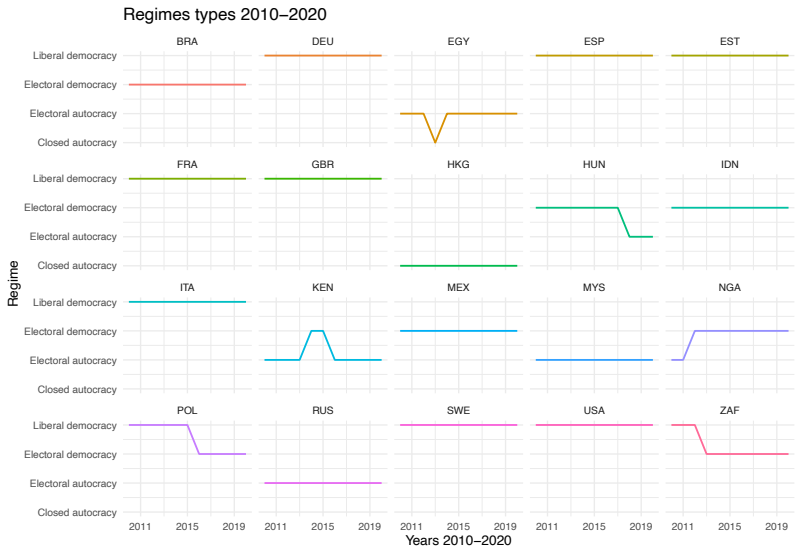
22. Liberal Democracy Index 2010-2020



Note: Higher index scores indicate higher degree of liberal democracy.

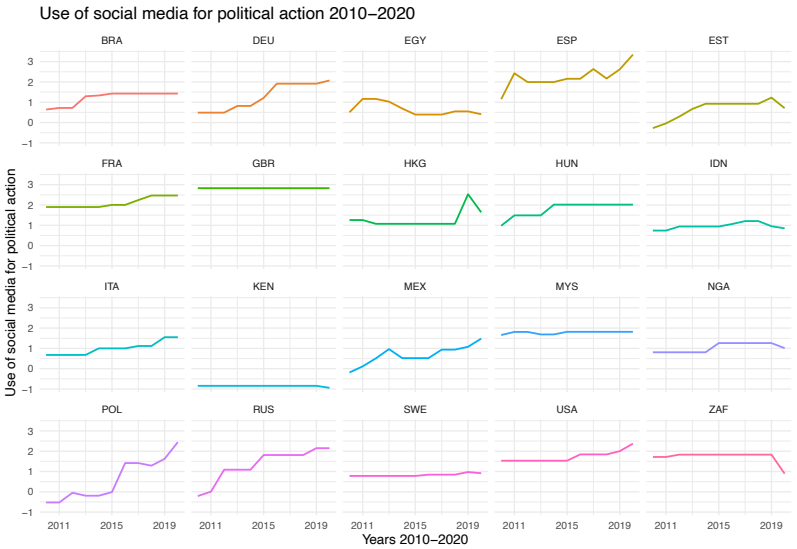
Graphs 22–31 come from the **Varieties of Democracy Project Time-Series Dataset**; Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, Nazifa Alizada, David Altman, Michael Bernhard, Agnes Cornell, M. Steven Fish, Lisa Gastaldi, Haakon Gjerløw, Adam Glynn, Allen Hicken, Garry Hindle, Nina Ilchenko, Joshua Krusell, Anna Lührmann, Seraphine F. Maerz, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Juraj Medzihorsky, Pamela Paxton, Daniel Pemstein, Josefine Pernes, Johannes von Römer, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Aksel Sundström, Eitan Tzelgov, Yi-ting Wang, Tore Wig, Steven Wilson and Daniel Ziblatt. 2021. V-Dem [Country–Year/Country–Date] Dataset v11.1. Varieties of Democracy (V-Dem) Project.

23. Regimes types 2010-2020



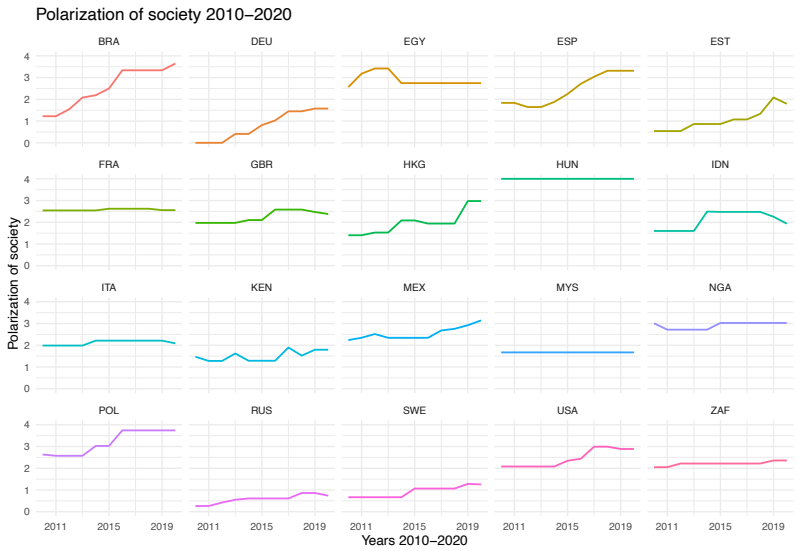
Original source: Lührmann, Tannenberg & Lindberg (2018); Varieties of Democracy Project.

24. Use of social media to organize offline political action 2010-2020



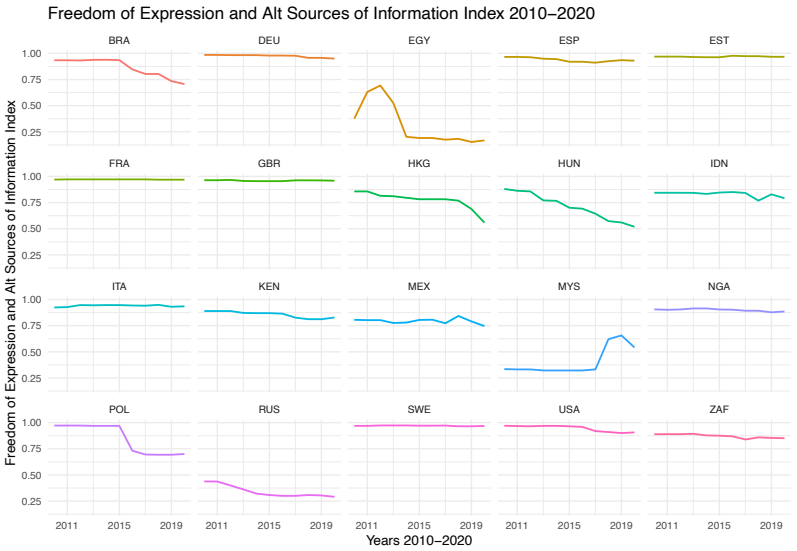
Original source: Digitial Society Project. Note: Higher scores indicate more frequent use of social media by people to organize offline political action of any kind.

25. Polarization of society 2010-2020



Original source: Digital Society Project. Note: Higher scores indicate greater polarization, i.e. differences in opinions on key political issues in society.

26. Freedom of Expression and Alternative Sources of Information Index 2010-2020



Note: Higher index scores indicate a higher degree of freedom of expression and alternative sources of information.

The Digital Society project paints a telling image of the countries where media operate under significant political control. Using the Liberal Democracy Index, we can see that a low score on it clearly corresponds to the likelihood of arrests

for counter-governmental content online (Graph 19). Please note that the index measures formal degrees of liberal democracy, i.e., the extent to which a country has achieved constitutionally protected civil liberties, strong rule of law, an independent judiciary and effective checks and balances that limit the exercise of executive power, and the level of electoral democracy. Out of our 20 countries, the countries that are found with the lowest LDI scores and the highest likelihood for arrests are Egypt, Russia, and Malaysia, soon followed by Nigeria and Kenya. The circumstances for social organization of political movements in these countries vary, however. With some preconceived ideas of how freedom of speech is affected under autocratic rule, the figures on Russia's polarization of political opinions among the general public suggest that the reason for complete uniformity is a stifled population aware that dissent leads to severe consequences (Graph 8, *Polarization of society*). Meanwhile in Egypt, polarization is at its highest, just as in Nigeria. Both Egypt and Russia still rank among the highest in the various forms of government control and surveillance of the internet and social media (Graphs 3–5), and Nigeria ranks fifth.

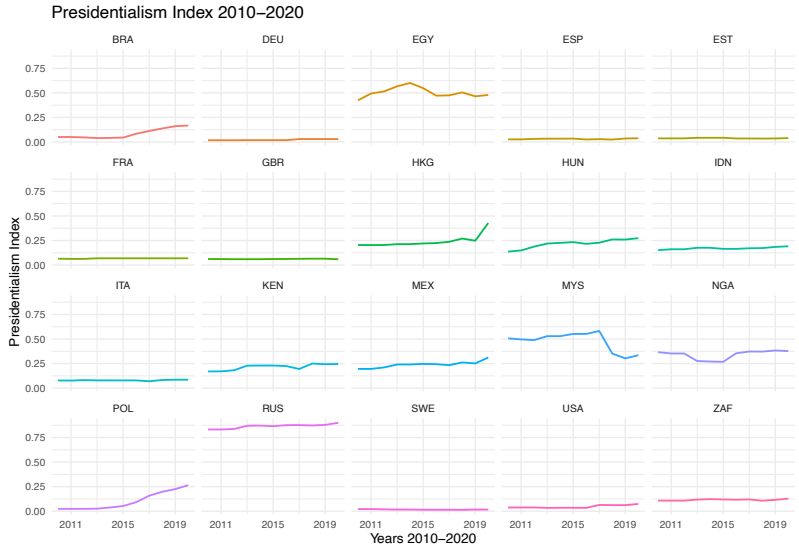
Regarding the countries that rank the opposite here, we see Hungary, France, Mexico, Italy, Spain, Germany, Great Britain, USA, Estonia, Poland, and Sweden among those with the least amount of government internet filtering and internet shut down (Graph 4). Apart from Hungary and Poland, these countries all rank among the highest in Liberal Democracy Index. Worth noting is that Hungary places the highest in this group with a considerably higher score on internet filtering. At the same time, Hungary, France, Spain and Poland all are among the countries with the highest degree of polarization, indicating that even the more polarized a society is, some governments may still not attempt to place considerable measures against the organization of dissent. Yet, it is perhaps not too surprising that high polarization may still be evident in countries with high Liberal Democratic Index.

When looking at Graph 5 (*Government social media monitoring and use of social media to organize offline political action*), it is apparent that in Kenya and Egypt, people use social media to a much higher degree than other countries to organize offline political action although there is social media monitoring in place. In Russia, social media use for this purpose is clearly lower, similarly as in Hong Kong where there are severe repercussions for participation in oppositional political action. Kenya is an interesting case, as it seems that political candidates heavily use social media to promote their 'offline' campaigning, bringing supporters to rallies or meetings with people out in public (Ndavula 2020). Political action might in this case also mean campaign meetings and rallies, and not always anti-government protests. Then there have been eruptions of protests that were met by police violence, as after the controversial 2017 elections.

In Graph 6 (*Use of social media to organize offline political action versus degree of governmental Internet shutdown*), the general trend seems to be that for higher occurrence of government internet shutdown, the higher the social media use for political action seems to be anyway. Shutdowns may happen when actions and manifestation grow and attract attention, but the organization stages might go unnoticed and might not prompt shutdown at that point. Perhaps this indicates that social media remains an efficient tool, even though suppression might be the outcome, as in the case of Kenya.

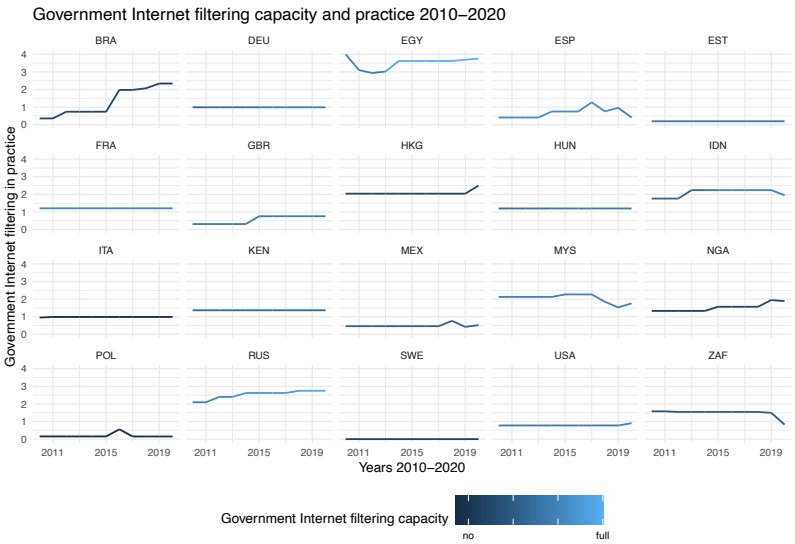
When it comes to media fractionalization (Graph 7), Hungary is at the top, even though the Fidesz government has strong ties with most news outlets and supports them financially. Spain's high fractionalization could perhaps be explained with the country's regional differences; autonomous regions with independence movements and where political rule subsidizes news outlets.

27. Presidentialism Index 2010-2020



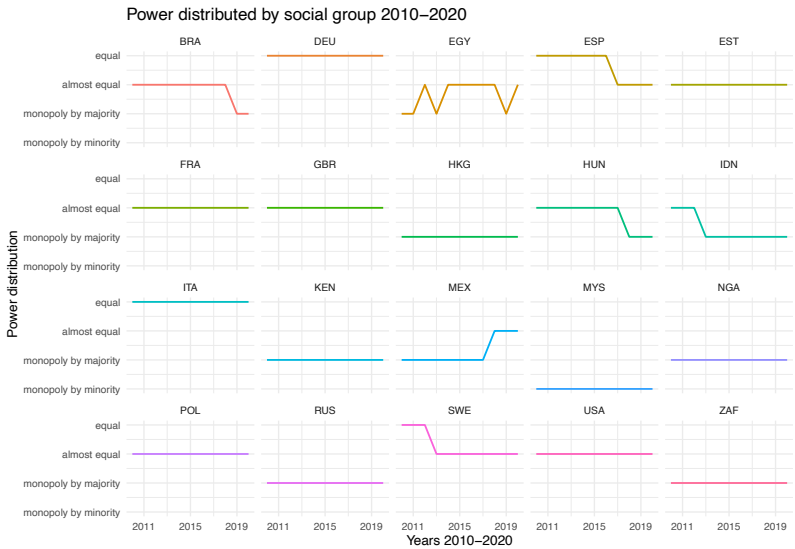
Original source: Sigman & Lindberg 2017; 2018; Pemstein et al. 2021; Varieties of Democracy Project. Note: Higher scores indicate higher degree of presidentialism, i.e. "systematic concentration of political power in the hands of one individual" (Bratton & Van de Walle 1997:63).

28. Government Internet filtering capacity and practice 2010-2020



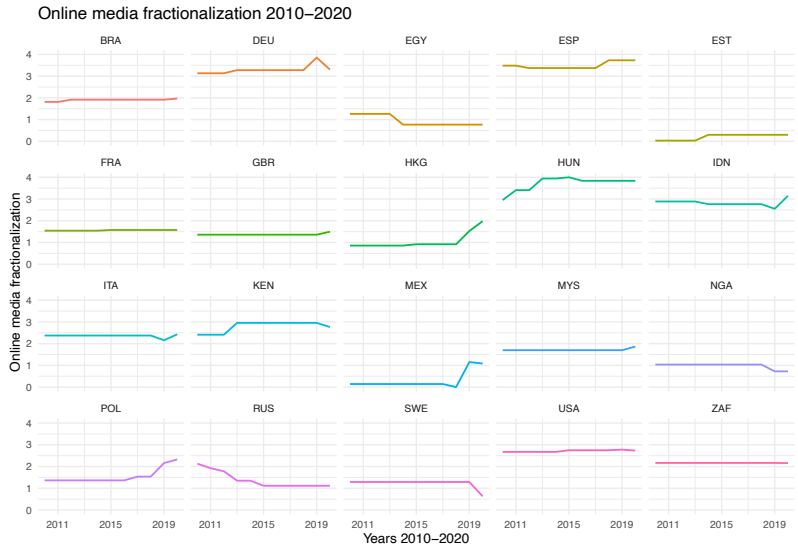
Original source: Digital Society Project. Note: Higer scores indicate higher government Internet filtering in practice.

30. Power distributed by social group 2010-2020



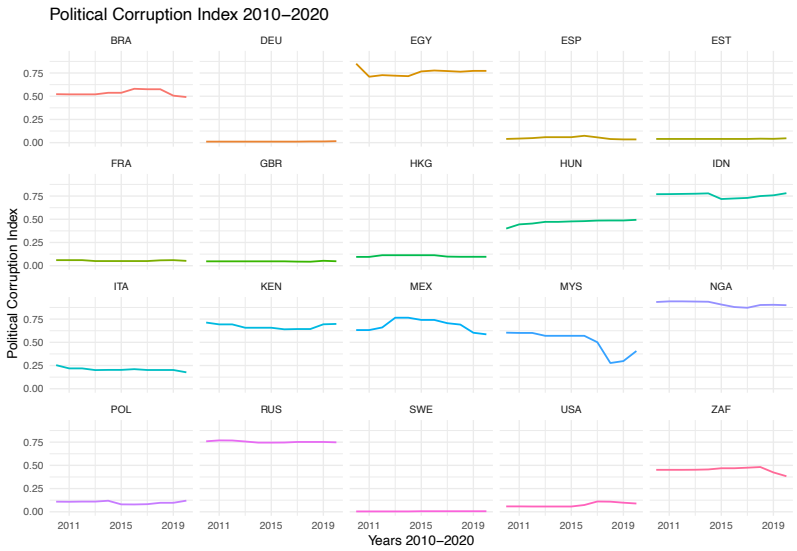
Note: Within each country, social group is differentiated on the basis of caste, ethnicity, language, race, religion, or other.

29. Online media fractionalization 2010-2020



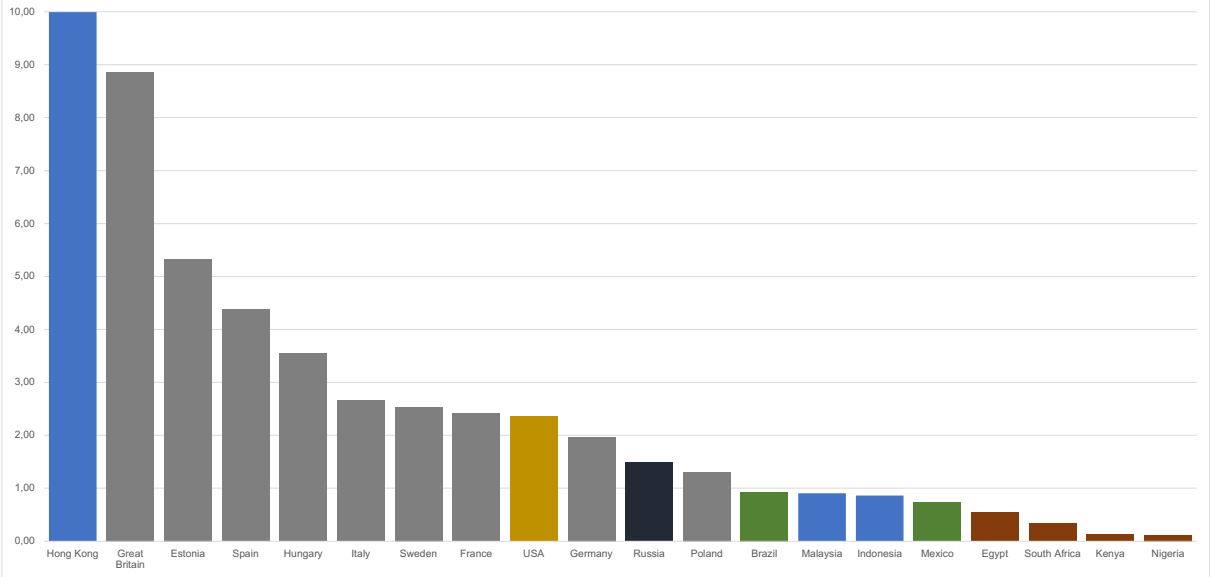
Original source: Digital Society Project. Note: Higer scores indicate greater polarization of key political issues in society.

31. Political Corruption Index 2010-2020



Note: Higher index scores indicate higher level of political corruption.

32. Penetration coefficients for online news media



Note: Cumulative traffic volume (SimilarWeb data) of the 10 most popular news websites per country, divided with total country population.
The coloring represents which part of the world each country is in (Eastern Asia, Europe, North America, South America, Africa, Russia).

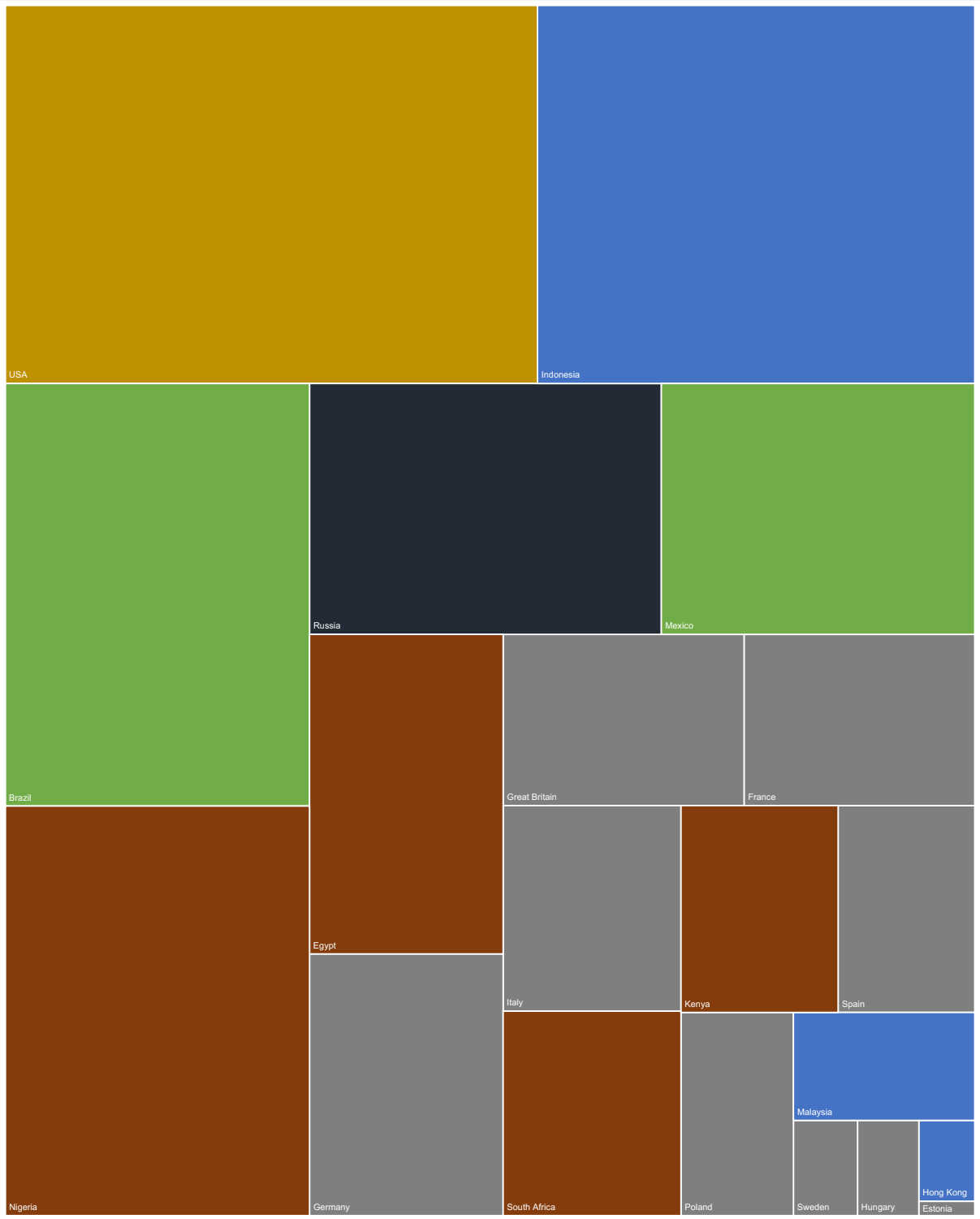
How has social media use shifted over time? Use of social media for political organization has been rising in almost all countries in the past decade (Graph 24). A few exceptions are Egypt, Hong Kong, South Africa, Kenya and Nigeria. In Egypt this peaked at the time of the 2011 revolution, when protestors mobilized through social media platforms. But since then, the country has become more of an autocracy – today Egypt monitors social media and arrests political opponents of the regime. Among the countries where perceptions of corruption are high (Hungary, Brazil, Indonesia, Russia, Mexico and Nigeria, Egypt has the least amount of social media organization. This could of course be related to the fact that less than half the population use the internet (Graph 12), and the country's telecommunications infrastructure is poorly developed as well (Graph 13). South Africa has seen the sharpest drop, and this has happened in 2019–2020. This could have related to the intense protests in 2017 having declined after Jacob Zuma's 2018 resignation from office and the period prior to his trial. It is likely however that the index for 2021 will show an increase again, with the turmoil and deadly protests erupting in July. Social media allegedly played a large part in the incitement for violence on the part of pro-Zuma protestors and rioters and was also used to spread false information. This saw the result of increased surveillance of social media, aided by the legislation of the country's cybersecurity act of 2020 (Karombo 2021).

Finally, the graphs show that in Germany, Spain, Poland, Russia – and to some degree Mexico and the US –social media use has grown in importance for political activism. To some extent this could be explained by the growth of far/alt-right movements, which in recent years of the Covid-19 pandemic have begun protesting restrictions and vaccinations.

There are some interesting correspondences between the QOG data and the SimilarWeb traffic data. E.g., the QOG data indicates high fractionalization of media in Hungary and Spain, whereas the SimilarWeb traffic data also confirms that these countries have comparatively popular online media titles. A reasonable interpretation of these indicators is that fractionalization in sheer quantitative terms is neither good nor bad, but rather an indicator that there is a breadth of media discourse on offer, in the supply side of online news media in the country in question. Likewise, the QOG data shows that Russia has a very low degree of polarization whereas, e.g., Sweden has low scores for polarization as well. However, this is likely to be due to entirely different reasons: In Russia, apparent media polarization is low since oppositional media are, quite simply, censored and forced out of the country, while in Sweden, apparent media polarization is likely low because of a political culture of relative homogeneity and compromise.

Which countries, in our overview, can be seen as "most online," when it comes to news websites? Given the various provisos noted above, and modestly accounting for the possible quirks of Web traffic estimates, we can still note some patterns that are both convincing and reasonable. As our metrics give indicative data for total data traffic for the various text-based news sites in question, these figures can be correlated with the individual population sizes in each respective country, to gauge a measure of how large the online news titles are in relation to overall population (Graph 32). When doing so, we found ratios for each title listed, and by compiling all these ratios for the top ten sites for each country, one would arguably get a standardized measure of aggregated impact for online news sites in each country. By doing so, we saw a huge variety in this weighted index value, ranging from 9.99 for Hong Kong (a figure that should be approached very carefully, as it largely results from the huge audiences in the neighboring countries for some of the media titles nominally belonging to the Hong Kong dataset, skewing the metrics in this way) and 8.86 for Great Britain (this figure is partially due to the outsized transnational audiences for leading online news sites BBC and The Guardian, nominally British but having enormous global web traffic) – to the African countries in our selection, all ranking at the very bottom of this comparative visualization. In this comparison, Nigeria's weighted online news presence, relative to population, is a mere 75th of Britain's.

33. Comparative population sizes of countries



Note: Total country population, represented as area (for visual comparison).
The coloring represents which part of the world each country is in (Eastern Asia, Europe, North America, South America, Africa, Russia).

Brazil

Population: approx. 212.6 million (2021)

Comparative digital news media penetration coefficient (our estimate): **0.93**

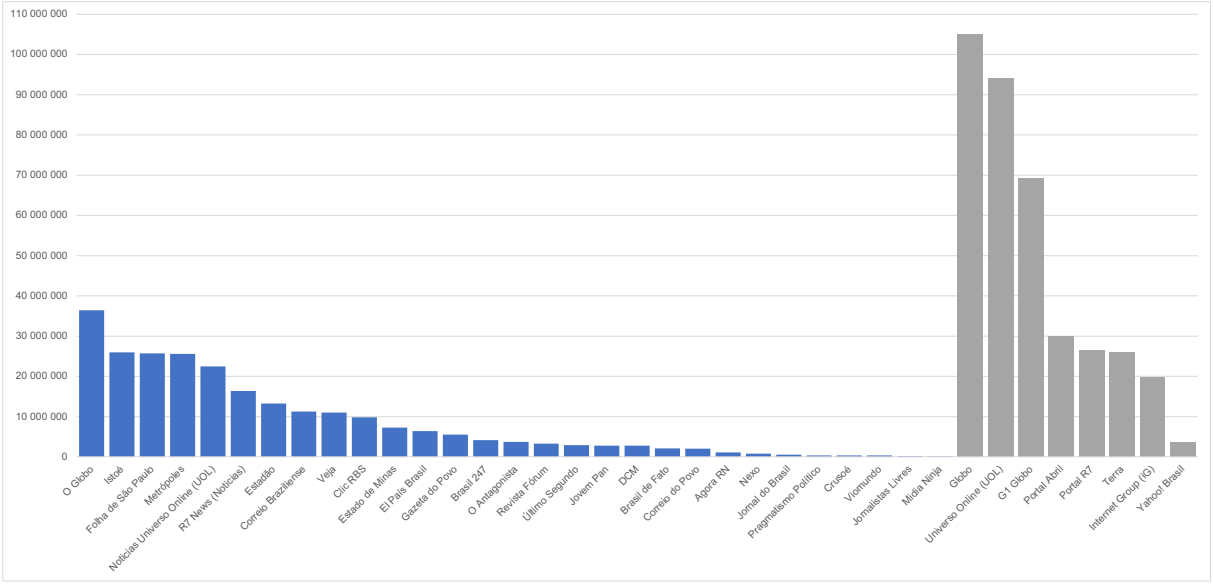
The role of presidents is crucial in the Brazilian political system. From 2003 to December 31st, 2010, the country was run by Luiz Inácio Lula da Silva, followed by Dilma Rousseff (2011–2016) from the leftist Worker's Party (PT). In 2015 she began her second term, but in 2016 the Brazilian Congress removed her from office by impeachment, being succeeded by her vice-president Michel Temer from right MDB (Brazilian Democratic Movement). In 2018, extreme-right Jair Bolsonaro (PSL – Social Liberal Party) was elected, taking office on January 1st, 2019.

Lula da Silva's presidency was accompanied by oil discoveries, continuous economic growth, and social reforms that ensured political support for his re-election in 2006. The country's constitution does not allow a president to run for a third term, and Lula da Silva was succeeded by his former Minister of Energy and later Chief of Staff Dilma Rousseff (Britannica 2021). However, Rousseff's presidency as the first woman president was far from triumphant. Rousseff's popularity fluctuated significantly, beginning at high levels, but dropping drastically from 2013 and onwards, in a turbulent political landscape characterized by protests. The demonstrations in 2013 began as a protest against the increase in the price of bus tickets, but later shifted into addressing other issues like corruption and the country's hosting of the football World Cup. In June 2013, many demonstrations were held across the country against the poor quality of social services and the expenses put out for the upcoming World Cup. In 2015, a major corruption scandal revealed links between state officials and the state-owned oil company Petrobras, in what became known as the "Carwash Operation" (Operação Lava-Jato) whereby many top politicians were indicted. This operation garnered a popular support and – combined with accusations against Rousseff for technicalities regarding the state budget – resulted in her impeachment and removal from office in May 2016. Many of the scandalous kickbacks were said to have happened during the years when Rousseff was the chairwoman of Petrobras, from 2003 to 2010. No direct evidence implicating Rousseff in the scheme has however been made public, and she denies having any prior knowledge of it.

The impeachment also entailed the end of a long stretch of Worker's Party (PT) rule of 13 years and a comparatively economically prosperous era (Romero 2016). Although this event once again illustrated the deeply rooted corruption in the Brazilian system, Rousseff's supporters view her removal as a threat to democracy (Romero 2016). She was succeeded by her vice-president Michel Temer, who served as interim president until 2018. Notably, after 2015, our macro level data (Graph 27) indicates that governance became less liberal with strengthened presidential powers. Contextually, it is associated with president's Dilma Rousseff removal from office by the Congress and the election of Jair Bolsonaro in 2019. In

2019, Jair Bolsonaro began his four-year term exalting the role of the president against a system of checks and balances prescribed in the Brazilian constitution of 1988, with a political vision to reform political institutions by undermining the power of the Supreme Court and appointing several military executives to top offices with large popular support despite repeated protests against his government's draconian measures regarding culture, education and civil rights. Bolsonaro's three sons are elected politicians; Flavio is a congressman in the state of Rio de Janeiro, Eduardo is a senator for São Paulo and Carlos is a member of the Rio de Janeiro City Council. In 2020 and 2021 the country experienced coronavirus protests over the slow response to the pandemic outbreak, causing Bolsonaro's popularity to plummet in view of the world's second highest death toll at more than 600,000 deaths. Parliamentary investigations regarding electoral manipulation using fake news, and corruption scandals involving his sons, who are all elected officers in state and federal constituencies, contributes to Bolsonaro's decline in popularity.

34. Brazil's most popular news/editorial websites (left) and editorial aggregators / web portals (right) according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



According to Media Ownership Monitor (2020), Brazil's media landscape is characterized by high concentration of audience and ownership, high geographic concentration, lack of transparency, and, moreover, religious, political, and economic interference. Brazil ranks poorly in the Press Freedom Index, at place 111 out of 180. In their most recent report, Reporters Without Borders don't mince words:

> Brazil continues to be an especially violent country for the media, and many journalists have been killed in connection with their work. In most cases, these reporters, radio hosts, bloggers or other kinds of information providers were covering stories linked to corruption, public policy or organised crime in small or mid-sized cities, where they are more vulnerable. Journalism has become especially problematic since Jair Bolsonaro's election as president in October 2018. Insulting, denigrating, stigmatising and humiliating journalists has become President Bolsonaro's trademark. (Reporters Without Borders 2021)

The biggest online outlets (portals that offer news and entertainment) are owned by private media conglomerates that already have a strong footing in print and television: Globo, UOL, Abril, and ClicRBS. In their review of internet media in Brazil, Valente & Pita (2018: 15) note that only 9 out of the 100 most accessed websites in Brazil were journalistic, and half of these were related to the nationally dominant Globo and Folha media groups.

In terms of media, Brazil is perhaps most known for its television market. The phenomenon of has been well-known for decades and is synonymous with syndicated soap operas in the world, mostly produced by TV Globo. Its owner, Grupo Globo, is the biggest media company in all of South America; a huge media conglomerate, founded in 1925, which owns magazines, newspapers and and Rede Globo, the second-largest commercial TV network in annual revenue worldwide behind just of American Broadcasting Company (ABC). The company operates G1, the largest news portal in Brazil, and continues to dominate print media with the national newspaper O Globo, and Editora Globo which publishes the weekly Época, the second largest in the country in terms of circulation.

Globo's main rival is Editora Abril, founded in 1950, which operates primarily in publishing, and owns a range of lifestyle-oriented media brands, including the largest news weekly news magazine, Revista Veja, way ahead of Época. Grupo Folha/UOL is another dominant media conglomerate ranking at number three in the country, a merger between three newspapers (Folha de São Paulo, Agora São Paulo, and Alô Negócios), internet provider and online portal Universo Online (UOL), and TV Folha (which is available through streaming on the UOL portal). Terra is a Spanish internet provider, part of telecoms giant Telefónica Group which has a significant foothold in Brazilian telecommunications since the privatization of the sector in the late 1990s. Its web portal terra.com.br was one of the first to emerge during the early 1990s and is still very popular in Brazil.

Paywalls are rather common as a means of financing online media in Brazil, restricting access to non-subscribers. At least a third of the 30 largest newspapers in Brazil have adopted paywall solutions, including Folha de São Paulo, O Globo, and Correio Braziliense (owned by Diários Associados, the 10th largest media conglomerate in Brazil).

In one sense, a similar argument can be made in Brazil as for many other countries in South America and Africa (see, e.g., Kenya), namely that radio has a significant impact, especially in the more rural areas. Of notable interest in recent years, during the Bolsonaro administration, is the growth of the Jovem Pan media conglomerate, a radio network that is highly expedient to the president's politics, indeed favored by Bolsonaro as a preferable channel for government-friendly communication, which largely exerts his populistic views and uses fake news often produced and circulated by the government itself.

But in another sense, Brazil has an altogether different media landscape compared to African countries of the same size (e.g., Nigeria, with the same size of population as Brazil), namely that in Brazil, there are some very large online media and telecoms brands – Globo, UOL, Abril, Terra, IG, R7. Much of the Web traffic is exclusively domestic, and not from diasporas in other overseas countries. In the listing of online media in Brazil, a phenomenon was noted where SimilarWeb has metrics for top domains (globo.com; uol.com; abril.com.br; R7.com; ig.com.br) but where individual publications can be found as subdomains. We listed both, as both metrics can be relevant as measures for comparison.

Out of our 20 countries, Brazil is not ideally compared with the Sub-Saharan African countries but should rather be compared with Mexico and Indonesia in terms of cultural-political impact in its region, government structure, media landscape, and the size and economic clout of its population. Brazil doesn't have a sizeable diaspora, but its 214 million inhabitants, with Portuguese as its dominant language, explains Globo's impact in Brazil and other Portuguese-speaking countries. In general, reading news is a middle-class phenomenon, however. The poor do not generally read news, but rather watch tv or enjoy social media, especially WhatsApp, to a large extent also Facebook and YouTube.

Like with Mexico, we have omitted those publications that are oriented almost exclusively to sports and celebrity/gossip/entertainment. Similarly, the distinction between pure-play broadcasting and text-based news is hard to make, as several of the big broadcasters also feature text-based news on their web portals.

Approximately 76–82 % of traffic comes from mobile devices. Some exceptions to this pattern are publications like online magazine Nexo, and online newspapers Jornal do Brasil, and Estadão, which see a comparatively more desktop-based, less mobile-based readership. This is due to the target audiences of these publications which are mostly urban and older middle-class readers, also given the low circulation of these outlets in comparison to, e.g., O Globo and others.

We have also endeavored to include explicitly leftist and progressive media, like Brasil 247, Revista Fórum, Midia Ninja, and The Intercept Brasil (the latter was

hard to track for web traffic, however, as it is hosted on the main US-based the-intercept.com domain), but traffic for many of these more alternative media outlets are vanishingly small, compared to the mainstream titles.

Egypt

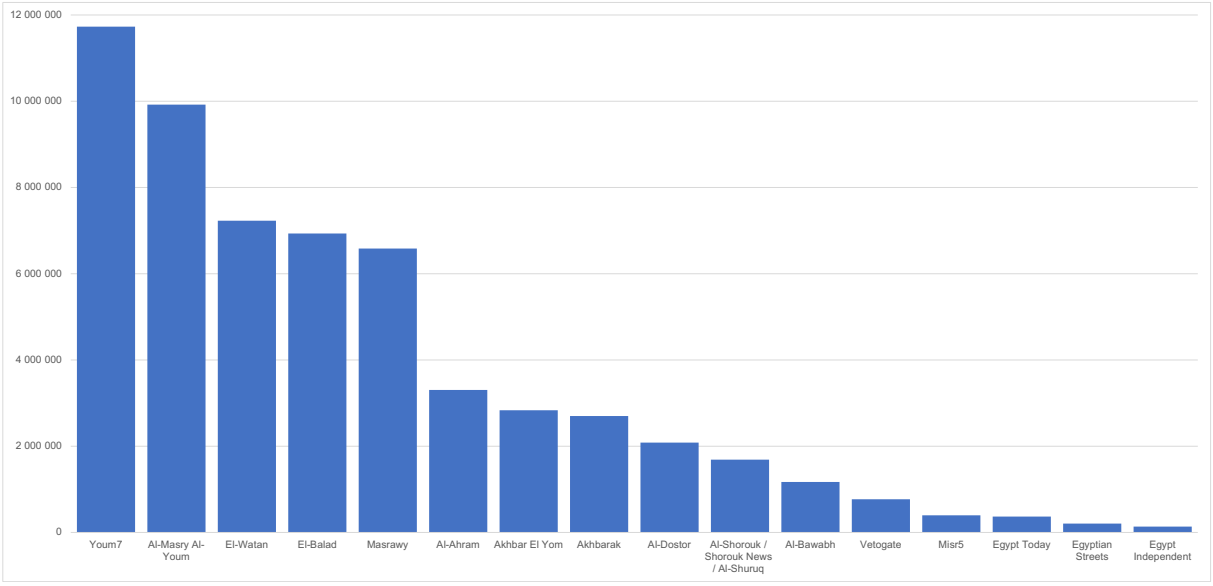
Population: approx. 102.4 million (2021)

Comparative digital news media penetration coefficient (our estimate): **0.54**

Egypt's 2010–2020 decade started with the revolution. Authoritarian leader Hosni Mubarak resigned after nationwide protests escalated against his almost 30-year rule and voter rigging at the elections. In 2012 Mohammed Morsi, a candidate from the Muslim Brotherhood party, won presidential elections. He attempted to approve a new constitution, restricting freedom of speech and expression. However, mass protests escalated in 2013, leading to the army overthrowing the president, declaration of the Muslim Brotherhood as a terrorist group, and banning religious parties in the newly approved constitution. In 2014 former army chief Abdul Fattah al-Sisi won the presidential elections and began to accumulate power. After this time, the enhanced role of the president, deterioration of freedoms, and Internet filtering by the government can be observed (Freedom House 2021).

In April 2019, President Abdel-Fattah el-Sisi initiated a referendum which allowed him to stay in power util 2030, expanding the president's control over judicial appointments and elevating the role of military in society (Freedom House 2020). Alongside this, numerous terrorist attacks were perpetrated by the Islamic State in Egypt. Considering the shifting power between military, civilians, and religious parties, the shifts in power distribution in Egypt are notable during the period.

35. Egypt's most popular news/editorial websites according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



Being the largest Arab-speaking country (with an estimated 101 million inhabitants in 2021) editorial media and news reporting in Egypt effectively affect the entire pan-Arabic political agenda. Egypt has a large number of printed publications, mainly in Arabic but also in other languages such as English, French and Armenian. For the research purposes in this project, the main has been on Web editions, not paper editions. According to translator services consultancy Industry Arabic (2020), it is mainly Levantine and Gulf regions that dominate news cycles and media markets in the Arabic language. Egypt, they continue, has earned a negative reputation when it comes to press freedom in recent years. Many smaller, independent, and oppositional newspapers have witnessed reprisals from the government, including raids, arrests, and forced shutdowns.

Mobile traffic constitutes 86–94 % of the traffic to the most popular sites. Looking at the graphs, it is clear that Youm7 is the market leader, in terms of Web traffic. Youm7 (Arabic: اليوم السابع, The Seventh Day) is a privately owned daily, founded in 2008, with a fairly independent editorial footing. It is arguably the leading Egyptian news outlet with broad pan-Arabic reach, especially among younger, college-educated populations. Forbes Middle East has named it "the most effective news website in the Middle East" (Industry Arabic 2020). Quite a lot of traffic (around 8 % for a site like Youm7) comes from Saudi Arabia, and around 3 % of traffic seemingly from the US. However, this latter figure can be partially due to internet users masking their Web surfing by directing it through VPNs and proxies.

According to Industry Arabic (2020), Akhbar el-Yom (Arabic: أخبار اليوم, News of the Day or Today's News), founded in 1944, has a very strong pan-Arabic reach, but according to our figures its domestic Web presence, in terms of raw traffic, is fairly modest compared to its competitors. It is the semi-official media arm of the Egyptian Shura Council, which gives it a privileged position within the country's media landscape. Instead, Al-Masry Al-Youm (Arabic: المصري اليوم, The Egyptian Today), founded in 2004, turns out to be the second-biggest online news publication after Youm7. It is, similarly, privately owned daily with an independent/reformist/liberal agenda. The third position in terms of traffic is shared by El-Watan, El-Balad, and Masrawy. Notably, another (very) traditional newspaper, Al-Ahram (الأهرام, The Pyramids), a legacy media institution in Egypt since its inception in 1875, while being the most widely circulating Egyptian daily newspaper, does not rank above the five more popular online news outlets mentioned above. It is owned by the Egyptian government via the Al-Ahram Publishing House, and logically very pro-government.

One of the rare oppositional newspapers in Egypt is Mada Masr (مدى مصر), which was raided by plainclothes police in November 2019.

> Mada Masr publishes reports about corruption and security issues in a manner that is often critical of the government and has become one of

> hundreds of websites that have been blocked by Egyptian authorities in recent years. […] More journalists are jailed in Egypt than in any country other than China and Turkey, according to the Committee to Protect Journalists, a watchdog based in New York. (Deutsche Welle 2019)

Despite being blocked in Egypt since 2017, the Mada Masr website is accessible in the country via virtual private networks (VPNs). Its traffic of is very marginal, its total page visits per month only just over 100,000 according to SimilarWeb. Out of those, only around 20 % nominally comes from Egypt, but the use of VPNs is common for knowledgeable media users in the country. A social medium that is often used for dissemination of online news that are critical of the Egyptian government is Twitter.

Estonia

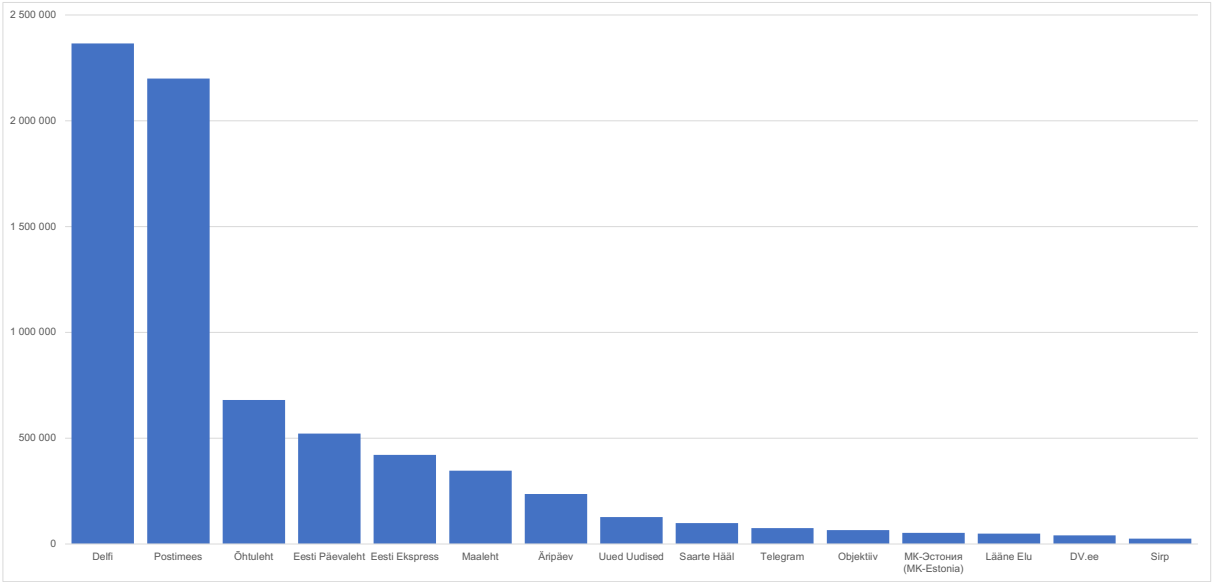
Population: approx. 1.33 million (2021)

Comparative digital news media penetration coefficient (our estimate): **5.33**

After the collapse of the Soviet Union in 1991 and the subsequent independence of many former Soviet republics, Estonia consolidated its autonomy and domestic development and decided to join the EU in 2004. In 2011, the country also joined the Eurozone and adopted the Euro as its currency.

From 2010 to 2020, Estonia has been governed by six different cabinets. In the 2011 elections, the Reform party, together with coalition partner IRL, recaptured the parliamentary majority, while Toomas Hendrik Ilves was re-elected as president. Andrus Ansip from the Reform party acted as prime minister in the 2011–2014 period (Casal Bértoa 2021). He was succeeded by co-partisan Taavi Rõivas (prime minister in 2014–2016). The Reform party won the 2015 parliamentary elections, accompanied by intensified tensions with Russia during the Ukrainian crisis. However, at the 2015 elections, the Reform and Centre parties had relatively close results (30 and 27 seats in the parliament, respectively), and in 2016 the country's prime minister became the Centre candidate Jüri Ratas, as Rõivas was taken down with vote of confidence in the end of 2016 after a sexual harassment scandal. After the 2019 election, which was won by the Reform party which however failed to build the necessary coalition, his Centre party infamously managed to build a coalition with the far right/anti-immigration EKRE party and the conservative Isamaa party. In early 2021, Ratas had to resign after taking political responsibility for a scandal of suspected illegal influence peddling, and Kaja Kallas from the Reform party became the country's first female prime minister.

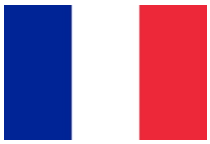
36. Estonia's most popular news/editorial websites according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



Estonia ranks as number 15 in the World Press Freedom Index and is often held up as one of the most advanced digital societies in the world, offering online provision of all sorts of governmental services and occupying second place for "internet freedom" in the Freedom House Index 2021. Around half of the Estonian voters use the Internet to vote in the national elections and European Parliament elections. Despite these positive trends, Estonia ranks as an "almost equal society" in indexes of power distribution by group indicator. This can be related to the fact that 6 % of the country's population remains stateless, without the right to vote in the national elections. Ethnic Russians, Roma, and other groups face discrimination, while corruption remains a challenge (Freedom House 2020).

In terms of frequency of visits, there are interesting patterns. Web portal and news aggregator Delfi is the website with most page visits in the country. With 12 page visits per unique visitor, it also stands out from the others in a significant way (see more on "stickiness" in section 3.3.2 above). It is important to note that Delfi is big in Latvia and Lithuania too, which could explain its extraordinary traffic volume. Included in our list we also have Delfi's sub-sites Eesti Päevaleht (daily newspaper), Eesti Ekspress (weekly paper) and Maaleht (Estonia's biggest weekly paper) which all have significant traffic as well. While all of Estonia's major newspapers have online editions, Delfi is the online-only news portal that has the most extensive readership and exists in several languages and editions. Estonia's biggest conventional online newspaper, without comparison, is Postimees. With an estimated 2.2 million unique visitors per month, there are few newspapers in any other countries in the world which even come close in terms of popularity in relation to population size. Its competitor Õhtuleht has almost 700,000.

For such a small country like Estonia, Delfi is humongous, with almost 2.4 million unique visitors per month. Its Russian edition alone seems to have around a million unique visitors. This observation, however, is a good illustration of how, as a metric, "monthly unique visitors" is *not* isomorphic with real human beings. Delfi and Postimees have more monthly unique visitors than there are people in Estonia! (Since over 90 % of these monthly unique visitors are from Estonia, according to SimilarWeb's own metrics, it is simply illogical to believe that "unique visitors" would always translate into actual human beings.) This could be explained by the fact that Estonia is a highly digitalized country, where over 75 % of the population has access to mobile internet, and where almost 92 % has internet access in their own household (Statistics Estonia 2021). This could have the effect that one unique visitor in fact shows up as two or three in the statistics due to multiple devices. Nevertheless, most internet users in Estonia tend to visit Delfi regularly.



France

Population: approx. 65.3 million (2021)

Comparative digital news media penetration coefficient (our estimate): **2.41**

France has one of the most polarized party systems in the EU, meaning that the general ideological distance between the parties is considerable, and anti-political-establishment parties capture a significant share of votes in the elections (Casal Bértoa 2021). In France, as in Italy, the rotation of cabinets is also persistent. From 2010 to 2020, there were 14 governments run by five different prime ministers (ibid.). Since the French political system is semi-presidential, the role of the president as the head of the state is also important. Nicolas Sarkozy served as the president from 2007 to 2012 and was succeeded by François Hollande, who acted as president until 2017. In the 2017 presidential elections, far-right candidate Marine Le Pen lost to centrist Emmanuel Macron, the current French president. His movement "La Republique en Marche!" also gained a majority in the parliamentary elections.

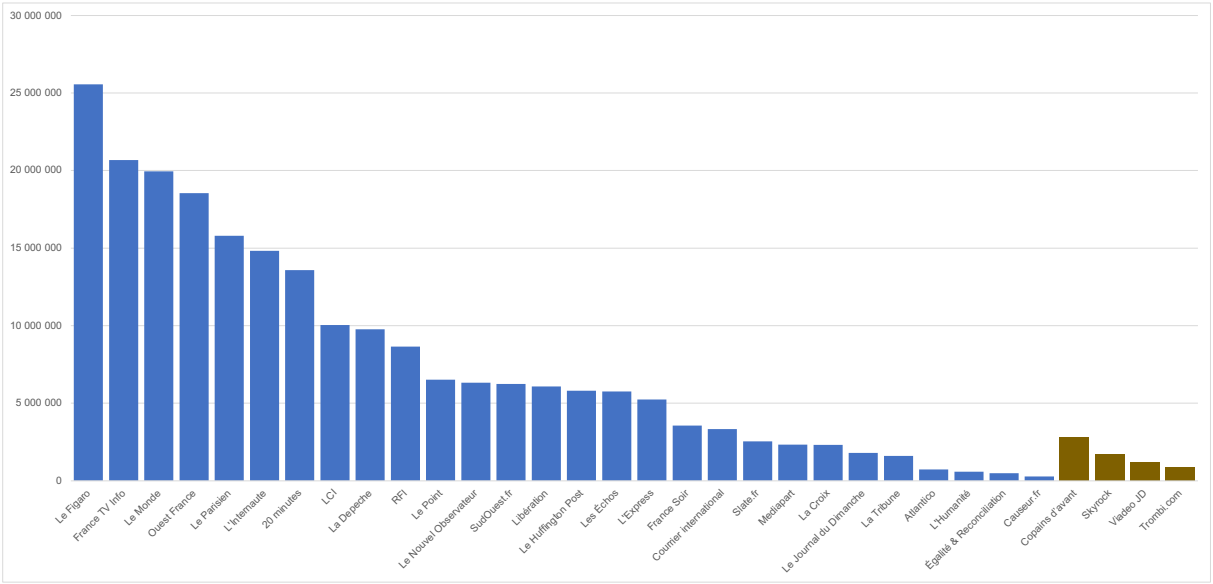
The consequences of the 2008 financial crisis were harmful in France, leading to budget cuts. In 2010, waves of protests were held in the country against government plans to raise the retirement age to 62. The political landscape is characterized also by military intervention in Mali (in 2013), several terrorist attacks (notable ones in 2015 and 2016), and subsequent anti-Islamist laws.

Nevertheless, the indexes charting political regime, political polarization, corruption levels, and media fractionalization remain constant across 2010–2020. A slight increase in media usage for protesting can be spotted. Besides, the French government filters content on the Internet to a greater degree than governments in Spain, Sweden, and Poland. On the face of it, polarization might seem to have increased: The far-right National Front saw a gain in votes in the 2014 elections, and in 2018 nationwide "yellow vest" protests were held in the country by railway workers against labor market reforms. But as in Spain and Italy, the party system in France has a historical legacy of being highly polarized, and anti-establishment parties such as the far-right National Front and the far-left parties Workers' Struggle, Revolutionary Communist League, and Extreme Gauche receive plenty of support. New parties have also appeared in the political landscape, such as far-left parties Parti de Gauche and Nouveau Parti Anticapitaliste.

Like many other European countries, France has a very strong history of newspaper culture, which is apparent when observing the internet media in France, since the online news sphere is clearly dominated by conventional legacy newspapers, often with strong heritage and legitimacy. This is arguably reflected in our graph, as it shows a rather plentiful variety of media titles, 17 of them with over

five million monthly unique visitors. The phenomenon on online-only news aggregators, common in many of the other countries in our brief overview, does not seem to have caught on in France. However, there are a couple of internet-born publications – e.g., L'internaute, founded in the year 2000.

37. France's most popular news/editorial websites (left) and online forums / social websites (right) according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



Looking at web traffic data, the French market seems somewhat tiered. Top news publications like L'internaute share the top tier of audience reach together with a handful other ones (Le Figaro, Le Monde, and a few others), all with a reach of over 15 million unique visits per month. Then, there is a second tier of popular news media, where around ten publications reach 5–10 million unique visitors per month, and a third tier of 1–5 million unique visitors per month.

The share of mobile versus desktop traffic to sites tends to be at around 25–33 % desktop-based traffic, versus 66–75 % mobile-based. Regarding transnational audiences, for most of the popular websites their audiences are largely domestic French audiences. RFI has got a significant readership in China, the US, and Canada.

During the early years of the millennium, there were attempts in the French-speaking online world to establish domestic social-media platforms. However, like in most other European countries, that market has been almost entirely captured by American tech corporations like Facebook, Google, Snapchat, and Twitter. In recent internet history, we can note several examples of domestic social

media in France, that have emerged but no longer has a strong footing: Le-BonFil.com, launched by a small French company in 2015, emerged and then disappeared without making much of a trace. Other notable examples: Skyblogs, a blogging site launched by Skyrock radio in 2002 where users could create personal pages; Viadeo, a professional networking platform in the style of LinkedIn, founded in 2004 and taken over by Le Figaro in 2016; Copains d'avant, a social networking website founded in 2001, allowing former classmates to stay in touch with each other, later bought up by L'Internaute and now largely insignificant compared to Facebook.

Germany

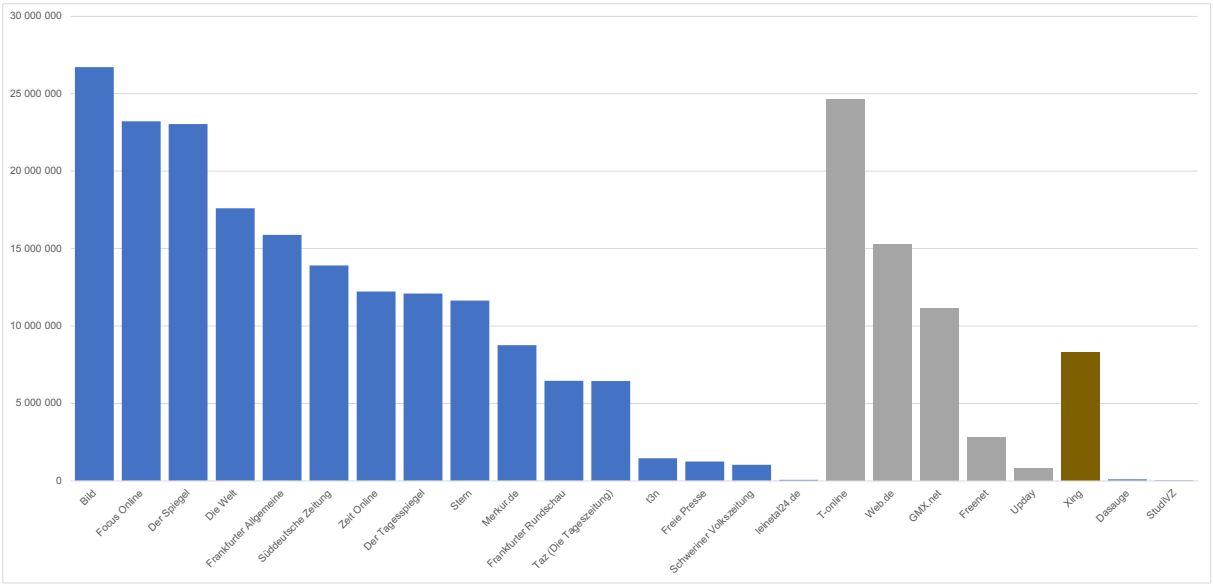
Population: approx. 83.8 million (2021)

Comparative digital news media penetration coefficient (our estimate): **1.97**

Angela Merkel was the chancellor of Germany from 2005 to 2021 (succeeded by Olaf Scholz). During her time in government, the country saw seen several monumental challenges. The management of the Eurozone crisis (2009–2014) heavily depended on Germany. For instance, Chancellor Merkel insisted on a second bailout for Greece to protect the Euro. During the migration crisis in 2015 and Merkel's third term of office as chancellor, the country provided asylum for refugees escaping the war in Syria and tried to combat the crisis. While immigration to the county had been continuously growing, it reached a peak during the years around 2015; the country had allowed approximately 800,000–900,000 asylum seekers in 2015 but restricted this number to 280,000 in the following year. There was a deal made between the EU and Turkey, and the routes from the Balkans were closed down. The influence of populist authoritarianism and opposition to this openness to asylum seekers prompted the rise of right-wing populism and strong results in some parts of the federation for the xenophobic Alternative for Germany party. This party entered parliament for the first time in the 2017 federal elections. Despite the rise of the far-right, political polarization within the party system is relatively modest in Germany (Casal Bértoa 2021).

From 2010 onwards, there has been a significant increase of polarization and, moreover, increased rates of use of social media to organize political action. All other indicators in our data stay relatively constant.

38. Germany's most popular news/editorial websites (blue), editorial aggregators / web portals (gray), and online forums / social websites (brown) according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



Germany is home to several large transnational media corporations, being dominant especially in markets for broadcasting and publishing: Bertelsmann, Axel Springer SE, ProSiebenSat.1 Media, and Bauer Media Group. Germany's market for newspapers and magazines is the largest in Europe. Legacy newspapers are often relatively young and fairly liberal – notably, almost all of the popular newspapers were founded during the postwar de-Nazification.

Regarding online news provision, however, when going through the various domestic news sites it is apparent that online news provision in Germany is not as clearly a remediation of paper editions as the online news providers are in Italy and France. In Germany, news portals (T-online, GMX.net, Web.de in particular) are as popular as singular online newspapers. Also, broadcasters such as N-tv and Tagesschau offer text-based news, competing with the legacy newspapers' online editions.

Looking at our graph, the most popular online news brands for text-based news are Bild, Focus Online, and Der Spiegel. The variety of news brands is not fully as wide as in France or in Britain, something that might be explained by a higher degree of regionality in the German press system. Around 60–70 % of traffic to news sites comes from mobile devices.

Great Britain

Population: approx. 68.3 million (2021)

Comparative digital news media penetration coefficient (our estimate): **8.86**

For Britain, there are relatively stable trends for all analysis indicators included – except for polarization, where the values are slightly higher since 2015. This trend might be related to such significant events as 2015 victory of the Conservative Party in the general elections (for the first time since 1992), and the 2016 referendum on leaving the EU. In the post-Brexit landscape, we have seen Prime Minister David Cameron's resignation, Theresa May's time in office, and, subsequently, from 2019 onwards, Boris Johnson's.

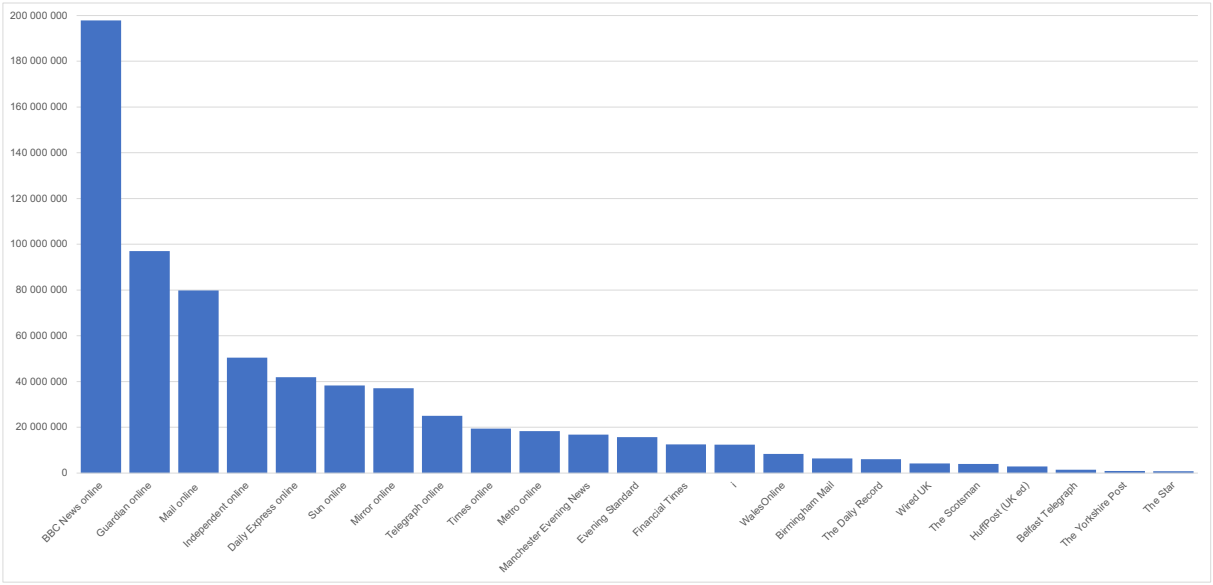
Notably, in the UK, the average use of media for civic protest remains constant for the whole period. The reason may be that the indicator was initially rather high, to begin with, at the beginning of the 2010s. Overall, the country has been ruled by the six cabinets during 2010–2020. Two of Cameron's conservative governments acted from 2010 to 2016. The other two conservative cabinets, ruled by May between 2016 and 2019, were short-lived. She was succeeded by the current prime minister Johnson in 2019.

Like some of the other peripheral countries in the EU, Great Britain has a rather different pattern of civic trust in the media compared to the main tendency in the EU. Great Britain is an outlier in indexes of public trust in the media, especially when it comes to the press. There is usually a pronounced distrust of the press among the British, clearly noticeable in Eurobarometer data (Strömbäck 2021). In 2021, public trust in the press soared at over 40 % but usually it lies around 20 % in Great Britain, far lower than its neighboring countries like the Netherlands (where trust in the press remains at around 70 %) or Denmark and Sweden (60 %). The low levels of trust in the press held by the British are more like patterns found in Cyprus or Bulgaria. Likewise, radio holds equally low levels of public trust in the UK, similar to Greece and Spain. Television, however, holds a slightly higher degree of public trust among the British; the French, the Spaniards, and, most notably, the Greek hold lower levels of trust in television (Strömbäck 2021).

Also, in terms of the relationship between general political trust and trust in the press, Great Britain holds an outlier position together with Russia, Greece, Serbia, Malta, and Moldova (Ariely 2015). Usually, trust in the press and political trust tends to go together, but in countries in which the media is more restricted and journalists are less professional, there is a stronger than average relation between trust in the press and political trust. In countries with a partisan newspaper bias, there is also a stronger relation (Ariely 2015). Britain seems to be an exponent of the latter category; it has a nominally free and powerful press, albeit

with very self-serving editorial positions due to a long tradition of partisan newspapers. In more recent, qualitatively oriented research, Palmer et al. (2020) have shown that the perception that newspapers fail their watchdog role and make part of a distant and self-serving political and economic establishment might be a rather common view among British citizens. Like with the discursive culture of disagreement notable in Hungary and the USA (see those sections in this report, p. 139 and p. 173), it is important to note what Ariely (2015) argues; credulous trust in politics is not necessarily a great ideal, since a critical view of politics is vital for democratic debate and for citizens holding politicians accountable.

39. Britain's most popular news/editorial websites according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



A methodological challenge when it comes to the English-language countries is that media that are nominally based in one country, as regards their top-domain URLs, can be accessed, and might indeed have significant audiences in entirely other countries. Take the UK Guardian newspaper, for example; it's estimated that a significant share of its online audience resides outside of the UK. The circulation of the paper edition of The Guardian is a mere 105,000 average daily sales (about half of which are subscribers, half paid single copies) with a clear downward trend for their paper edition, as average daily sales totaled around 187,000 in 2013 (according to ABC media auditing, June 2013) and have dropped by almost 44 % since then. However, the Guardian have paying supporters that range at around two million unique individuals, and when the evasive concept of "reach" (PAMCo 2018, n.d.) is estimated by The Guardian's advertising office,

they claim that the paper's monthly print reach is 3 million, while they claim to have a monthly cross-platform readership in the UK of 24.3 million, and an online reach of up towards 199 million people worldwide (The Guardian 2020). Note that digital reach is calculated through adding all visibly related content for a publisher brand, in various environments, including measures of daily, weekly and monthly "unique visitors" by platform, which, as we have seen above, is also an evasive measure. The British joint industry media measurement operator PAMCo (successor to the National Readership Survey) revised its methods for estimating reach in 2018, seen as there was a strong need to incorporate phone, tablet, desktop and print platforms for publishers (Tan 2018).

Both the Guardian and the BBC have .com addresses. However, whether the nominal top-domain URL is .com or .co.uk is of little relevance; the important factor is the above consideration of how large a share of the overall Web traffic metrics we display below are actually stemming from the target country (in this case, the UK). Moreover, American media brands like CNN, Buzzfeed, and MSN also have considerable audiences in the UK – operating as subdomain under these brands' .com domains, these editions are however hard to measure reliably in our traffic data. Conversely, some media conglomerates have top domains pertaining to one particular country but run specific national editions as sub-domains on the same website. Take the BBC, for example; they even run editions in different languages than English, catering for minority languages in former colonies like Ghana and Nigeria. The BBC, Financial Times, the Guardian and the Independent are notably transnational in their reach (BBC online has a lot of traffic coming from countries like the US, Germany, Canada, while the Guardian is particularly big also in Australia). The Independent seems to have a notable US audience. Broadcasters like Sky, ITV, and Channel 4 are not included in our listing – except the BBC, as its online edition has a strong legacy of written news.

For the popular websites, around 57–72 % of traffic comes from mobile devices. For a lot of titles (Independent, The Sun, The Mirror, Metro, iNews, The Express, The Times, Evening Standard) the share of mobile traffic is as high as 76–85 %. For Financial Times, on the other hand, the share of traffic that comes from mobile is only 46 %.

Hong Kong

Population: approx. 7.5 million (2021)

Comparative digital news media penetration coefficient (our estimate): **9.99**

During the period under scrutiny, the political leadership of Hong Kong has changed three times. In the early 2000s, the Chief Executive was Donald Tsang, who governed for two terms, followed by Leung Chun-Ying and Carrie Lam, who has been the incumbent since 2017.

The pro-Beijing administration criticized the first term of Donald Tsang for conducting structural economic reforms and exacerbating confrontations between rich and poor (Cheng 2007: 17). Hong Kong's economic competitiveness declined in 2008–2009 (p. 18). Besides, during Tsang's second term, big business and government conflicts intensified (p. 21). Although Tsang's governance was characterized by bureaucratic rule in the mainland's favor, he managed to cooperate with the Democratic Party to pass the electoral reform of 2020 on selecting the Chief executive and forming the Legislative Committee, which ensured a higher degree of decision-making for the territory (pp. 25–26).

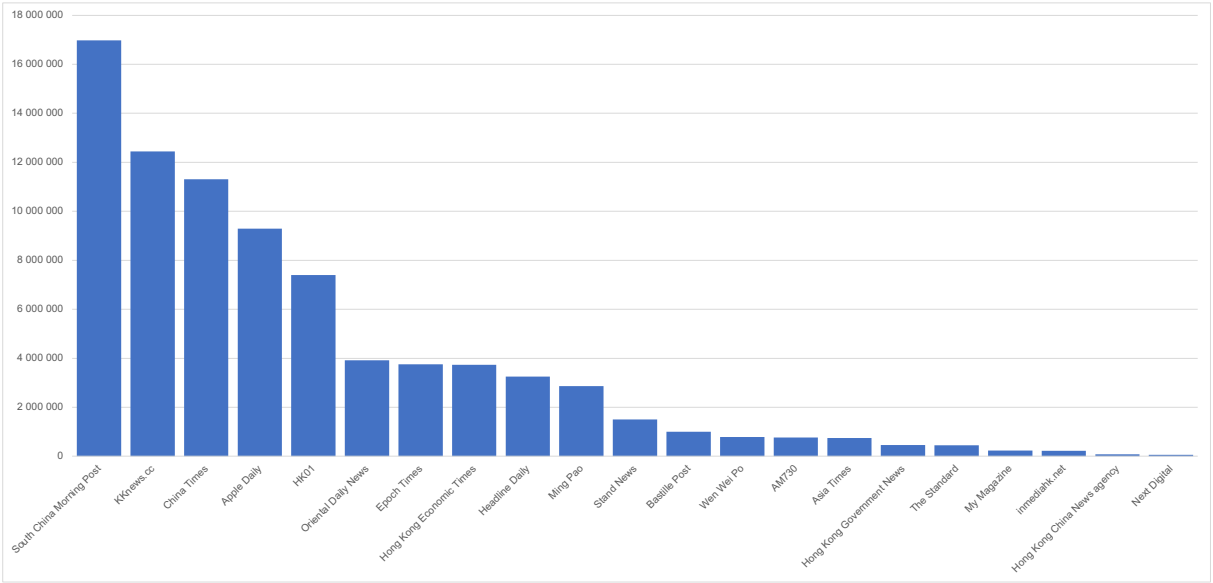
In 2012, Leung Chun-Ying stepped into office after corruption accusations and misconduct attributed to his predecessor Tsang. The new leader took office under circumstances of rising inequality and democratic demands. He attempted to introduce pro-China "patriotic lessons" in schools, indicating a significant political affiliation with Chinese politics, and decided to allow only those candidates nominated by the national committee to run in elections for the post of Chief Executive. Public dissatisfaction with this decision escalated in the so-called Umbrella Revolution, which took place in 2014. Students were protesting electoral reforms showing very high levels of public participation. The mass protests resulted in the withdrawal of the electoral proposal in 2017.

Since 2017, Carrie Lam has been occupying the post as Chief Executive, being the first female leader in this position. Lam is favored by Beijing, and her administration supported the controversial extradition bill of 2019. Another series of protests escalated in 2019–2020 against this law, passed by the government, which allows for the extradition of citizens to China and Taiwan. Protesters succeeded and the bill was retracted. In more recent years, tactical uses of social media to organize political actions have surged, and young protesters have invented new methods of mobilization, many of these involving the use of digital media.

Most Chinese speakers in Hong Kong speak the Cantonese dialect. While Mandarin is the official dialect of China and is used throughout the country for government communication, only around 48 percent of the Hong Kong population actually speak Mandarin (which is a significantly higher percentage than it used to

be – in 1996 it stood at only around 25 percent). However, Cantonese and Mandarin do use the same Chinese characters, albeit the whole of mainland China (excluding Taiwan, Hong Kong and overseas Chinese communities) tends to use simplified Chinese characters while Hong Kong, Taiwan, and Singapore have continued to use traditional script. Furthermore, in Hong Kong, although using traditional characters, its usage is different from Taiwan because in Taiwan the language used is Mandarin while in Hong Kong it is Cantonese. It is therefore challenging for Chinese people in mainland China, Hong Kong and Taiwan, respectively, to read each other's local media. One positive aspect of this, nevertheless, is that based on the languages used on different websites, we can infer their main audience. Moreover, while English is another official language in Hong Kong, around half of the Hong Kong population are fluent in English (this number has also increased over the last decades; in 1996, the figure was 38 percent). All official signs and announcements are in both Cantonese and English, and all government officials are required to comprehend English.

40. Hong Kong's most popular news/editorial websites according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



The vicinity to mainland China and the wide dispersion of Chinese as a language makes our traffic data figures difficult to interpret; many of the popular media titles in Hong Kong also have considerable audiences in mainland China and Taiwan, which makes for very large volumes of web traffic per title. Since it is so hard to break this traffic down into which respective jurisdiction traffic comes from, our overall global traffic volumes indicate monumental traffic, especially

for the news aggregators (Sohu.com and 360doc.com) and online forums (Tencent QQ, Zhihu) popular in Hong Kong. Traffic volumes for these titles dwarf the conventional news outlets' traffic: Sohu.com has 76 million average monthly unique visitors, 360.doc has 13 million, while QQ has a staggering 192 million and Zhihu 94 million. These popular websites are primarily Chinese; only a fraction of their traffic comes from Hong Kong.

Hong Kong has its own top domain, .hk, and the city state is home to the corporate headquarters of many prominent media houses, some of which have transnational reach (e.g., the Asian Wall Street Journal and Far Eastern Economic Review and, also, publications with anti-Communist legacies, e.g. The Epoch Times). In contrast to mainland China, where official control over the mass media is pervasive, Hong Kong houses several media outlets (both print and broadcasting) that are not subject to pro-Beijing censorship. Freedom of speech and of the press are enshrined in the Hong Kong Basic Law (the constitution of the Hong Kong Special Administrative Region).

In the mid-2010s, a lot of new media outlets cropped up in Hong Kong, and print newspapers were seriously challenged by a range of online-only news media. Briefly, a major reshuffle of the media ecosystem in Hong Kong was had taken place in 2016, as incumbent newspapers like The Hong Kong Sun folded that year while significant new online news portals like HK01, Bastille Post, Hong Kong Free Press, and Stand News were established. Around the same time, new online social forums like LIHKG and Zhihu were established.

A lot of online media are mostly reposting news or are gossip oriented. Sometimes the ownership structure is opaque, as offshore companies might hold ownership of the companies in question. With small scale and limited resources, online media primarily relies on traditional advertising. Therefore, in addition to news about current affairs, almost all websites also do a lot of entertainment and funny news to increase traffic.

Because online news is easy to reprint and plagiarize, the credibility of online media should be questioned (Wen 2016). Still, some of the online media outlets have been included in audience surveys on media credibility conducted by the Chinese University of Hong Kong (CUHK 2019). During the time of our investigation, the conflict between mainland China and Hong Kong loomed large – leading to concrete, very tangible events such as the closure of popular news outlet Apple Daily, in June 2021. Founded by famous Hong Kong entrepreneur and activist Jimmy Lai, it was one of the best-selling Chinese language newspapers in Hong Kong. On 17 June 2021, Hong Kong authorities froze the assets of the company and Lai personally. As a result, Apple Daily was forced to cease operations. The final print edition was printed in over a million copies (compared to the usual 80,000) and the closure was widely recognized internationally, indicative of

a significant blow for press freedom in the region. In our preliminary traffic data overviews in the summer of 2020, Apple Daily was the third most popular online news outlet of those listed for the region, with 9 million average unique visits per month (compared to Kknews with 18 million, and South China Morning Post with 14.7 million). A year later, in July and August 2021, Apple Daily was the fourth most popular, with 9 million average unique visits per month (compared to South China Morning Post with almost 17 million, kknews.cc with 12.4 million, and China Times with 11.3 million).

Between 65 and 80 % of traffic to the popular websites comes from mobile devices. In general, it's not strange that the traffic data from China and HK might be a little confusing. First, if people use VPN to visit a website, their location might show up as being in the USA while being in Hong Kong or elsewhere. It's very popular to use VPN in Hong Kong and China due to the significant government monitoring and restrictions. There are also many foreigners living in China and Hong Kong, who might be even more cautious in their online media behaviors. After the 2019–2020 Hong Kong protests, media restrictions in the region have arguably become more intrusive.

Secondly, South China Morning Post's URL scmp.com is its international edition, in English, and not intended for mainland China (in simplified Chinese). In China, even for singular media titles, it is common to separate a "domestic version" from an "overseas version." Therefore, Chinese people in China do not really read this version of South China Morning Post. So, it makes sense that 25 % of traffic is said to come from the US, 17 % from Hong Kong, 7 % from Singapore – and virtually none, or very little, for China.

Third, KKnews with the URL kknews.cc is a version in traditional Chinese characters (the common language script also in Hong Kong). Chinese people in mainland China would not browse this website since they use simplified Chinese characters. Two thirds of the documented traffic to KKnews comes from Taiwan, 15.5 % from Hong Kong, and 6.5 % from China.

Fourth, China Times (chinatimes.com) is a daily Chinese-language newspaper published in Taiwan, mainly reporting the news of Taiwan. It therefore makes sense that 71 % of its traffic was seen to come from Taiwan, 12 % from the US, 10 % from China and only 2.5 % from Hong Kong.

Lastly, the Hong Kong edition of Apple Daily ceased operations in June 2021. The version now available is the Taiwanese edition, so it makes sense that the numbers disclose that 87 % of its traffic is from Taiwan, and only 5 % from Hong Kong. In other words, in 2021 it is notable how much traffic Apple Daily seems to be getting from Taiwan, while Epoch Times gets a lot of its traffic from mainland China.

Hungary

Population: approx. 9.7 million (2021)

Comparative digital news media penetration coefficient (our estimate): **3.55**

Since the victory of the far-right party Fidesz in the parliamentary election of 2010, the country has been pushed in a clearly less liberal-democratic direction, with vastly increased rates of corruption. Democratic rights have now been continuously undermined in the country for a decade.

It's important to note that Hungary was in a dire economic situation in 2009, being forced to request to the IMF for loans. In combination with major dissent against the alleged foul play by the Hungarian socialist party MSZP under the incumbent prime minister Ferenc Gyurcsány (who had admitted to this in the so-called Őszöd speech) the Eurozone crisis did likely play a role when popular opinion turned against the Hungarian political left, which in turn precipitated the major victory of Fidesz, previously in the opposition, in the 2010 Hungarian parliamentary elections. Viktor Orbán has been the acting prime minister since then, forming four different cabinets over the last decade (Casal Bértoa 2021).

In 2011–2013 the government amended the constitution, electoral laws, and media laws, and also reformed the central bank legislature, facing criticism from the EU (Freedom House 2020). In the 2014 elections, Fidesz once again seized a supermajority in coalition with Christian-right party KDNP. It was the first election under the new Hungarian constitution which had come into force in 2012 – including new rules restricting the access and eligibility for citizens to vote. Moreover, by this time, also many of the major media outlets acted in ways beneficial to the incumbent Fidesz majority. In October 2014, after the elected government tried to pass a law that would tax Internet usage, peaceful protests emerged – possibly, the largest anti-government events since the protests in 2006 against the then-incumbent MSZP.

During the 2015 migration crisis, Hungary opposed the relocation of migrants and held a national referendum, where the government's stance was to support this position. The government pursued a strongly anti-immigrant rhetoric and passed laws such as that in 2017 that requires NGOs to register as foreign organizations if they receive funding from abroad. In the 2018 elections, once again the Fidesz–KDNP alliance emerged victorious. Popular protests emerged also after a so-called "slave law" was passed in 2018, which allows employers to delay payment and to force employees to work overtime.

Fidesz has openly tried to capture independent state institutions, compelling also nominally free enterprises to comply with the government agenda, enforcing greater government control over the court system, weakening independent institutions, exerting considerable political influence on the media and mounting an

increasingly hostile environment for civil-society organizations, along with corruption scandals involving EU funds. As the Covid-19 crisis emerged, the ruling majority adopted legal changes further centralizing its power and weakening its control of public spending, threatening to block the adoption of the EU budget and Covid-19 recovery funds. The country's illiberal trajectory has been emboldened by the government passing restrictions on academic freedom and actively campaigning against institutions that the government deems undesired, such as the Central European University.

Notably, Hungary is the most polarized society on major political issues in the studied sample (Graphs 8, 21, 25). Despite the deterioration of freedoms and the increased rate of presidentialism (Graph 27), the Hungarian government does not seem to try to filter the Internet, and the level of filtering on record is comparable with France and Germany. Compared to most other EU member states, corruption is very significant in Hungary and can be compared to African and Central-Asian countries rather than Western European. According to Transparency International, Hungary's administration has deteriorated from an already low score of 51 % in 2015 to 44 % in 2020, making Hungary the most corrupt EU member on par with Romania and Bulgaria.

Hungary is an interesting case study, partially because of this decline of deliberative democracy in the country, but partially also because Hungary has the highest rates of popular engagement with social media, according to Eurostat (Digital Economy and Society in the EU, 2017). Hungary had a domestic social-networking site called iWiW (launched in 2002), which was however eclipsed by the popularity of Facebook and other US-based international social media platforms. As of 2019, Facebook penetration in Hungary was estimated at around 62 % of the online population. The number of internet users was slightly over 89 % of the overall population. Judging from the SimilarWeb data, around 60–65 % of traffic to popular Hungarian news sites comes from mobile devices.

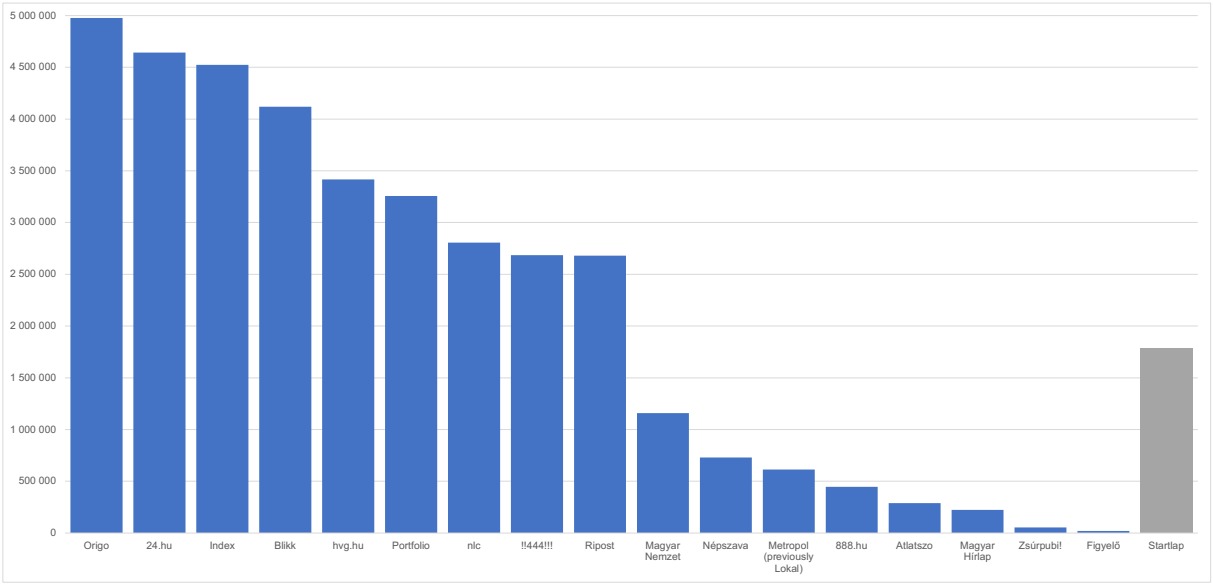
It has been established in the international community that the Fidesz government has attempted to obstruct much of the country's independent media (Hopkins 2021), something that is also reflected in Hungary's ranking in the World Press Freedom Index. Hungary now ranks on place 92 on that index, an abysmal position for a European country (however, on par with other countries further east and south, e.g., Serbia, Ukraine, Montenegro, Bulgaria, Russia, Belarus – all ranking even lower in the index). In 2020, the Fidesz government imposed a coronavirus-related law, with penalties of up to five years in prison for false information – "a completely disproportionate and coercive measure," according to Reporters Without Borders (2020).

Legacy media in Hungary have undergone a large-scale transformation, with an industrial structure of ownership and audience habits that are highly different

from those before the 2008–2009 economic crisis and the subsequent 2010 electoral victory of Orbán. It was after the 2014 election that the most significant changes came to pass, as the ruling coalition expanded its influence also over much of the private media, via a network of media oligarchs informally linked to the Fidesz party. Like in Italy during Berlusconi, the government controls both public service media and a significant share of commercial media and uses this power to coordinate propagandistic campaigns. The pro-government public service media receive more than 83.2 billion HUF (260 million EUR) per year, alongside "another few billions from the Media Council almost every year. In 2017, the value of state advertisements was 10 billion HUF […]. The distribution of public money is not in harmony with the relevant regulations of the European Commission" (Humán Platform 2020: 76). The pro-government media conglomerate relies not only on significant state funding but also on regulatory allowance to maintain a cartel. In February 2019, the 28 media companies operating under the umbrella of the Central European Press and Media Foundation (KESMA) entered into an agreement to form an officially recognized consortium: a media conglomerate involving up to 500 different media outlets, under the management of the Mediaworks holding company, which up until then had been owned by Hungary's wealthiest man, Lőrinc Mészáros, a close friend of Prime Minister Viktor Orbán and former mayor of Orbán's hometown. By gifting his entire media portfolio to the KESMA foundation, the merger could be legally exempted from competition law. According to a government decree by Orbán, this conglomerate constituted a "merger of strategic importance at a national level," legally protecting it from future state inquiries.

Freedom House (2018) has documented pro-government content, propaganda, and misinformation proliferating online in the lead-up to the 2018 parliamentary election. It has been noted that public officials have repeatedly initiated defamation campaigns and libel charges against citizens commenting on social networks.

41. Hungary's most popular news/editorial websites (left) and editorial aggregators / web portals (right) according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



The country has eight online news providers with highly loyal readers, forming a solid block of leading online media. Notably, the conservative news title Index tops our calculations of visits per unique visitors (which is a good measure of "stickiness" or audience loyalty; see section 3.3.2 above). The largest paper in terms of the sheer amount of monthly unique visitors, however, is Origo – previously, an investigative opposition publication, but as of recently transformed into a major pro-Fidesz outlet. In total, nine news sites have monthly readerships above two million, some of these are left/liberal leaning, such as 24.hu and 444. Compared with Sweden, which has a population of the same size and two news sites that stand out significantly, Hungarian online news might be argued to be not only reaching larger proportions of the national population, but more fractured in their political agendas and reporting and therefore precipitating a rather wide spread of audiences (see, e.g., Bajomi-Lazar n.d.).

When plotting the leading Hungarian editorial websites, it is apparent that the Fidesz-leaning tendency has become even more pronounced, only in the last year: In July 2020, the online news website Index, previously independent, succumbed to Fidesz pressure, as Indamedia (one of Index's important business partners) was purchased by Miklós Vaszily (a businessman with close ties to the Fidesz political party, and the president of TV2) in March 2020. On July 22, the Editor-in-chief of Index was ousted, and two days later there was a wave of resignations in the newspaper's head office, meaning that most of its editorial staff were replaced over the following months.

Prior to this, the country's other leading news website, Origo, had already turned pro-Fidesz (in 2014). Moreover, in October 2016, the country's largest political daily newspaper Népszabadság was suddenly closed (the newspaper's owner at the time was the abovementioned, Fidesz-affiliated company, Mediaworks). Observations such as these make for a strong case to argue that this is a highly biased news media landscape. Ripost and Magyar Nemzet are two other notably pro-Fidesz news publications.

Indonesia

Population: approx. 273.5 million (2021)

Comparative digital news media penetration coefficient (our estimate): **0.86**

By all estimates, Indonesia is one of the most linguistically diverse countries in the world, with perhaps 800 languages spoken, according to the 2010 census. The exact number of languages depends on whether to consider many of these to be dialects of the same language, but even a more conservative estimate is that around 700 languages would be spoken in this very populous island nation. The official language, however, is Indonesian (Bahasa Indonesia); a standardized version of Malay. Standard Indonesian is a formal, standardized language mainly designed for the written word, and arguably rarely reflective of the local vernacular (Fettling 2018).

During the 31-year rule by army officer and politician Suharto between 1967 and 1998, Indonesia was generally classified as a dictatorship, that came to leave a lasting legacy on governance and public administration in the country. After an amendment to the Indonesian constitution in 2002, the first direct presidential elections were held in 2004. Susilo Bambang Yudhoyono became the Indonesian president and ran the office until 2014. Yudhoyono, and the Democratic Party he represents, attempted to strengthen democracy, fight against corruption, and improve bureaucratic efficiency. Economic growth was solid. However, after the 2009 re-election, Yudhoyono's support started to erode, mainly to unsuccessful combat against corruption and nepotism. Scandals involving high public officials close to the president emerged.

Joko Widodo won the presidential elections in 2014 against an ex-general candidate. He was re-elected as president in 2019. Although he promised to break up with the authoritarian heritage, in practice, Widodo was engaged in compromises with corrupt politicians and religious leaders, tolerated human rights abuses and deterioration of rule of law (Bland 2019). Surrounded by generals, the president also reinforced the military role (ibid.)

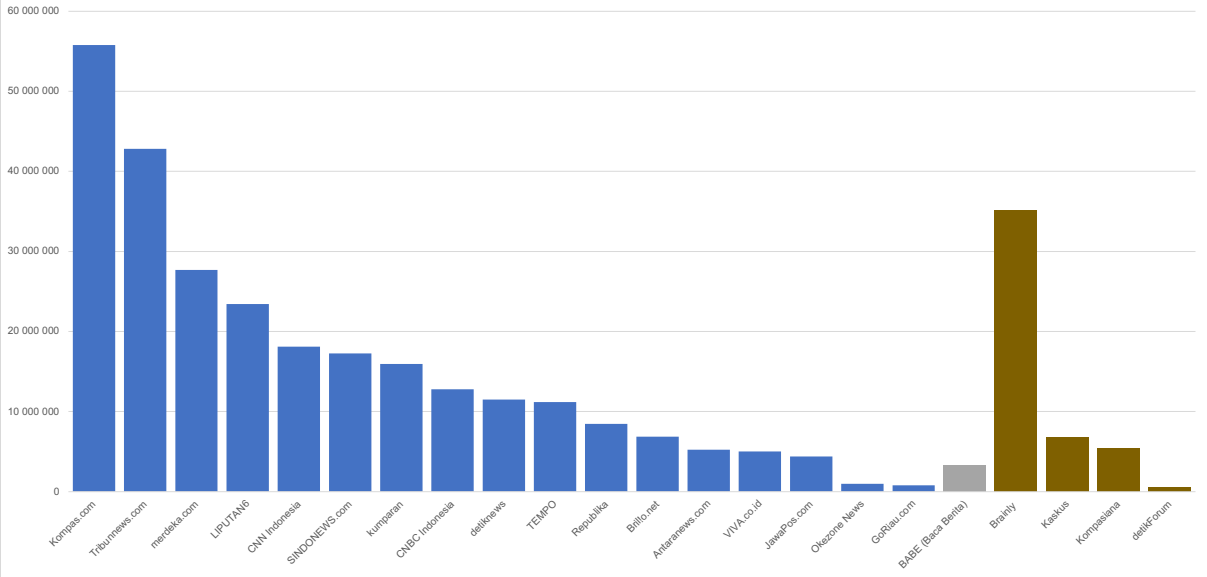
Variations of regime status over time have remained relatively stable in Indonesia. However, with a closer look, one can spot a slight drop on the liberal democracy index, an increase in societal polarization on major political issues, a slight increase in the president's power, and in filtering of the Internet by the government.

Indonesia ranks at place 113 in the 2021 Press Freedom Index. While far from perfect, press freedom improved considerably after the fall of Suharto's regime. At the time, Indonesia had also begun providing Internet access to citizens. The Indonesian press is sometimes said to be among the freest and liveliest in Asia, with a plethora of newspapers and magazines. The largest ones are Kompas

(Jakarta), Suara Merdeka (Semarang), Berita Buana (Jakarta), Pikiran Rakyat (Bandung), and Sinar Indonesia Baru (Medan). Large English-language dailies are Jakarta Post and Jakarta Globe.

Reporters Without Borders point out that there are nevertheless drastic restrictions on media access to West Papua (the Indonesian half of the island of New Guinea), where violence against local journalists keeps on growing, there are numerous abuses by the military, and cover-ups of humanitarian issues. The authorities also no longer hesitate to disconnect the Internet at times of tension. The Covid-19 crisis has allowed the government to reinforce its repression against journalists, who are now banned from publishing not only "false information" related to the coronavirus but also any "information hostile to the president or government" even if it is unrelated to the pandemic.

42. Indonesia's most popular news/editorial websites (blue), editorial aggregators / web portals (gray), and online forums / social websites (brown) according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



Over the last years, Indonesia has had the most spectacular internet penetration growth of any country in the world. In the mid-2010s, annual growth figures were at around 50 %. In absolute numbers, Indonesia comes fourth in the ranking of absolute numbers of Facebook users – only behind Brazil, India, and the United States. As of early 2020, Facebook penetration in Indonesia was estimated at around 80 % of the online population. The number of Internet users was slightly over 62 % of the total population. This digital progress is also mirrored in the tendencies seen in the traffic data: In Indonesia, as much as 90–97 % of the

traffic to news sites appears to come from mobile devices. The Reuters Institute Digital News Report (Newman et al. 2021) mentions that mobile aggregators is a rather new phenomenon that plays a significant role in many Asian markets, "partly due to bundling with local phone operators, partly due to the stronger penetration of Android devices, and partly because of a history of early mover advantage" (p. 27). Aggregators in Indonesia that Newman et al. (2021) note are: Line Today (20 % of the respondents stated that they have used it in the last week); Baca Berita (18 %); and LintasBerita (8 %). Out of these three aggregators, the latter's Indonesian URL (lintasberita.id) performs very poorly in traffic rank, while Baca Berita's traffic is similar to blog- and user-generated content aggregator Kompasiana and web forum Kaskus, but nowhere near as big as the traffic to the Indonesian edition of peer-to-peer e-learning website Brainly. In our traffic analysis, Line Today appears to be a pan-Asian aggregator, with substantial traffic but traffic coming from numerous countries (Taiwan, Thailand, Japan, Indonesia, Hong Kong) which makes it hard to conclusively pin down in individual country reports such as these. Kaskus is an online forum, with bloggers writing editorial content that is also open for comments. Kompasiana is a site where bloggers write editorial content. In terms of transnational traffic, Kaskus has significant overseas appeal; it attracts users from USA, Singapore, Germany, and a host of other countries.

Some US American broadcasters have a large footing in the online news ecosystem of Indonesia; CNN Indonesia as well as CNBC Indonesia are included in our listing, since they provide considerable text-based news as well as video footage.

The many Indonesian domains that we examined (most of them did not make the chart at all due to very small web traffic) were cross-checked with the source URLs used for news stories deemed valid in the manually annotated dataset by Rahutomo et al. (2018) as part of a study by Pratiwi et al. (2017).

Italy

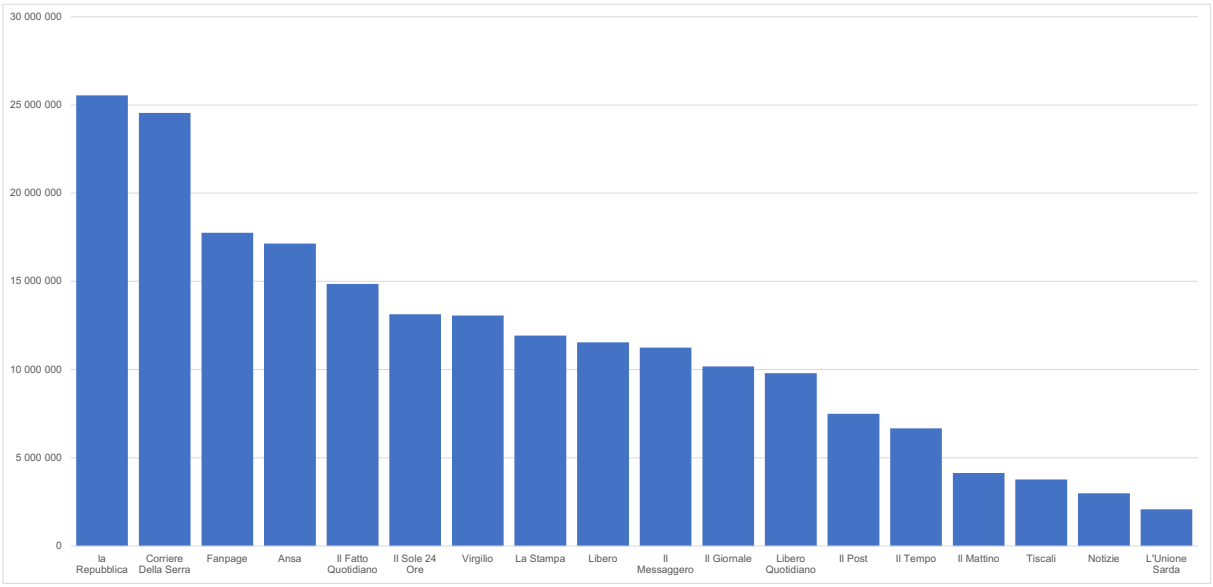
Population: approx. 60.5 million (2021)

Comparative digital news media penetration coefficient (our estimate): **2.66**

The most recent decade in Italy has been shaped perhaps most notably by the migration crisis, as well as corruption scandals (including former Prime Minister Silvio Berlusconi's sentencing for tax fraud), austerity measures following in the wake of the financial crisis in 2008–2009, and the rise of populism. Looking at key political indicators, Italy's positions have, however, remained the same. While the political corruption index measured by the V-Dem stays relatively constant across time, its values are higher for Italy, compared to other EU member states such as Spain and France. The use of social media to organize political actions appears more common in Italy, but the index values for Italy are still lower than in Poland, Hungary, and in many other European countries.

Italy is famous for its shifting parliamentary coalitions. During 2010–2020 there were 13 cabinets engaged in policymaking (Casal Bértoa 2021). Silvio Berlusconi was prime minister from 2008 to 2011, leading four different governments. In 2011 he was succeeded by the more technocratic leadership of Mario Monti. After him, Enrico Letta from the Democratic party attempted to create three short-lived cabinets after the 2013 general elections. Matteo Renzi, at the time secretary of the Democratic Party (PD), served as prime minister in 2014–2015, succeeded by co-partisan Paolo Gentiloni Silveri in 2016. General elections took place in 2018, and since those, three cabinets have been operational, managed by independent candidate Giuseppe Conte. After the Covid-19 pandemic, political crisis ensued and Mario Draghi, the former European Central Bank president, took office in February 2021. In general, while the party system is highly institutionalized in Italy, the ideological polarization between parties is very high – indeed one of the highest in all of the EU (Casal Bértoa 2021).

43. Italy's most popular news/editorial websites according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



A striking feature of the online news landscape in Italy is how the online editions of newspapers owe a lot of their design and market structure to the rich supply of print media in the country. La Repubblica, Corriere Della Serra, Il Sole 24 Ore, Il Messaggero, Il Fatto Quotidiano, La Stampa, etc. – all of these online newspapers are essentially print media, remediated into Web form. This means that the daily newspaper, as a typically 20th-century concept, seems to remain highly popular across Italy. From what we can glean from the metrics, around 70–85 % of traffic to these news sites comes from mobile devices.

As of 2019, the number of internet users was slightly over 92 % of the overall population, and Facebook penetration in Italy was estimated at around 55 % of the online population.

Kenya

Population: approx. 53.8 million (2021)

Comparative digital news media penetration coefficient (our estimate): **0.13**

During the 2002–2007 period, Mwai Kibaki served as president. In the presidential elections held in 2007, as Kibaki competed with the opposition candidate Odinga, political crisis ensued. The announcement of Kibaki's victory caused mass protests, claiming that elections were stolen (Brownsell 2013). More than 1000 people died in post-election violence, and ethnic conflicts were escalated (Associated Press 2013).

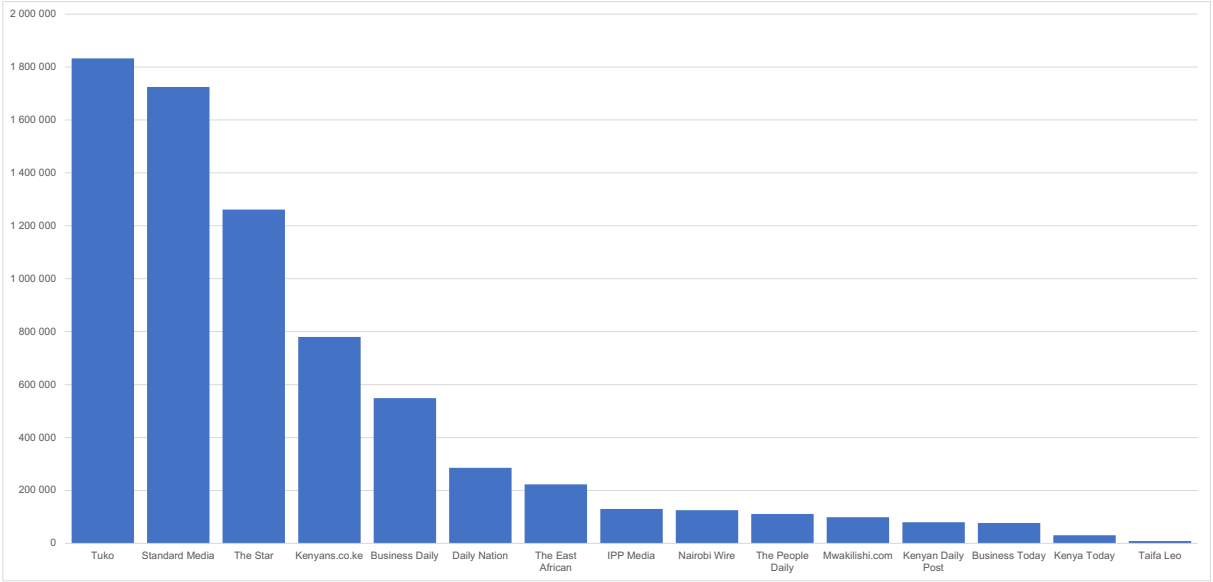
In 2010 a new constitution, aimed to limit the presidents' powers and devolve power to the regions, was adopted in the country after a national referendum. From 2012 to 2015, there was a peak of electoral democracy in Kenya, alongside oil discoveries in 2012, which brought a boom in the country's GDP growth. Still, authoritarian backlash emerged in 2015–2016. In 2013, Uhuru Kenyatta won the presidential elections and repeated his victory in 2017. However, the 2017 election results were highly contested, causing an eruption of protests and violence by police against protesters. The general elections were annulled, and the incumbent president obtained a victory at the polls, capturing 98 % of the vote. Besides, the period is characterized by terrorist attacks by Somali-based islamist group Al-Shabab in Nairobi in 2013.

Moreover, online news media appears to have become more fractionalized in the country during the observed period. It is instrumental to note that both president Uhuru Kenyatta and deputy president William Ruto have been owning considerable parts of broadcasting company Mediamax, and therefore exert power over the media in very direct ways. With digitalization, the media landscape has faced challenges, including disinformation, propaganda, and heavy influence from media giants in the advertising industry. Vital financing of the media comes from advertising in connection with current events such as political campaigns or sporting events that attract large audiences (Reelforge & TIFA 2019: 5–6). Kenyatta poses himself as a president open to international cooperation. However, during his presidency freedom of expression has been limited, and anti-corruption campaigns have not produced the desired results. Nevertheless, the president promotes the country's digitalization by ensuring better access to online government services and launching so-called e-centers.

According to a report written by Chaacha Mwita (2021) for the Kenyan office of international nonprofit organization Internews, social media are now the dominant source of information for many Kenyans, having surpassed radio as the primary source of news and entertainment – but this observation is likely only valid for urban populations. Africa is the least urbanized continent in the world, and one has to understand the differences between countryside and urban areas,

since these are gigantic: In Nairobi, most people are very connected to the internet, while in the countryside there are people (especially women) who have barely even watched TV and have never been online (despite having mobile phones that are technically equipped to browse the internet).

44. Kenya's most popular news/editorial websites according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



According to statistics from 2015, 98 % of the population, at that point, had access to radio, which is the dominant medium in the country, while mobile phones were becoming more common as, already in 2015, 97 % were estimated to have access to mobile phones. At the same time, when asked about having access at home or "elsewhere," 81 % claimed to have access to TV and 51 % to the internet. Note, however, that these are very spurious figures, since "elsewhere" can mean many things, especially in the rural context. The majority of all adults use radio and/or mobile phones daily and are the main sources of information on politics and news. Radio and television are also considered to be the most credible media for politics and news compared to newspapers and the internet. On the other hand, it is more likely for younger generations to be more influenced by internet and TV sources than older people, who show great confidence in radio and religious leaders such as priests and pastors (BBC Media Action 2018).

Since 2015, according to Reelforge & TIFA (2019), there has been a drastic development in what they call "digital migration," as most analog networks have been transformed into digital. Since 2015, the internet has become more accessible to

the population which has resulted in a reduced number of radio listeners – from 92 to 66 % in 2019. Mobile phones have also become more common, and access is estimated to be over 100 %, which means that a sizeable part of the adult population has several phones or SIM cards. However, radio is still considered to be the dominant and most credible medium in Kenya, as the total audiences are much larger compared to newspapers and TV (Reelforge & TIFA 2019: 10–18). Vernacular or community-based languages are mainly represented in radio, not in online text-based news.

Othieno Nyanjom (2012) has shown, in another Internews report, how mass media in countries like Kenya often serve as a nexus between those in power and the general public; "politicians have increasingly taken advantage of media liberalization to directly or indirectly acquire media interests with which to secure their place in politics. […] politicians have realized that acquiring votes through the airwaves is more cost-effective than traversing a constituency personally or using proxies, giving monetary or material favors" (p. 41). Many media companies might suffer from corrupt policies and lack of transparency, as journalism is manipulated by biased media owners or politicians vying to influence the agenda. In addition, most employees in the media industry are low-educated and low-income earners with poor employment conditions, which means that large parts of the Kenyan media landscape, according to Nyanjom, reflects a lack of professionalism. This is a problem that is most pronounced for radio and television, as these unidirectional mass media are vastly popular and reach almost the entire population.

In our statistics, it is notable that around 80–85 % of overall traffic is mobile traffic. The leading online news outlet, in terms of sheer traffic, is Tuko, a Kenyan online newspaper and entertainment website, established in 2015 and having gained in popularity very rapidly, mixing aggregated, exclusive and users' generated news content. A significant share of its traffic comes from other countries – especially the US, Canada, and South Africa. Its contenders, Standard Media and The Star, are leading legacy newspapers. As we can see, online news is entirely dominated by English-language outlets; a pattern similar to the other sub-Saharan African countries in our overview – Nigeria (in particular) but also South Africa. The main Swahili-language newspaper in Kenya, Taifa Leo, has tiny web traffic in comparison.

It should also be noted how comparatively small the traffic figures are for Kenyan online media; as our digital news media penetration value indicates, penetration rates are 0.13 which is comparable only to the other African countries in our overview. The figures indicate that African online news media still are very marginal in terms of Web traffic in comparison to population size, so much so that one could look at this media sphere as qualitatively different in societal impact and industrial size, compared to those of other continents.

Malaysia

Population: 152pprox.. 32.4 million (2021)

Comparative digital news media penetration coefficient (our estimate): **0.91**

Malaysia is a federal constitutional monarchy with a very strong Muslim hegemony, although its administrative and governmental tradition is secular and based on English Common Law, granting freedom of religion to non-Muslims and nominal rights to the various ethnic groups inhabiting the island nation. It is still a matter of public debate in the country, as to which degree Malaysia should follow secular or Islamic principles. Since 2018, the Malaysian government has overall become more liberal, less corrupted, seeming to better protect freedom of expression – but at the same time, parliamentary disorder has been considerable, culminating in a full-blown political crisis in 2020–2021.

Mohammad Najib Abdul Razak was the prime minister from 2009 until 2018. The so-called 1Malaysia Development Berhad scandal (1MDB scandal) in 2015–2016, directly involving prime minister Najib Razak, caused an eruption of protests, and his party lost the 2018 election and accepted the election results, promising to help facilitate a smooth transition of power. Despite economic liberalization reforms conducted during his time in office, the political opposition was silenced, and corruption and kleptocracy was thriving.

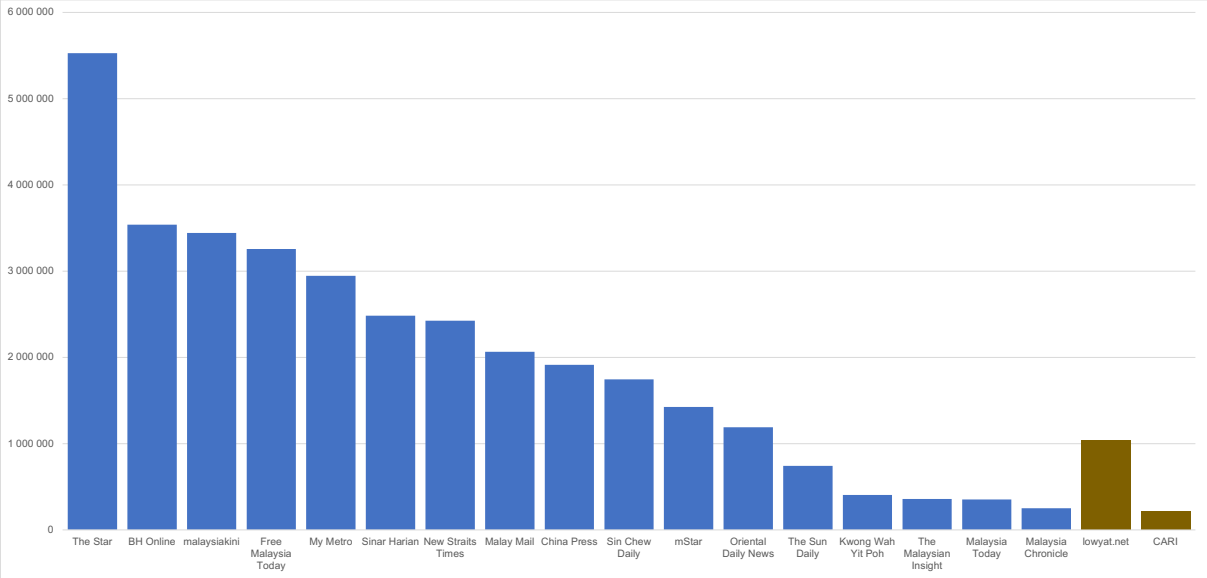
After the 2018 elections, Mahathir Mohamad became prime minister, leading a coalition of four parties. This has been argued to be a turning point, as the presidential powers could arguably be said to have weakened in the country. However, his government collapsed in 2020 and a new one was formed by Muhyiddin Yassin, with political instability continuing throughout 2020 and into 2021, exacerbated by the COVID-19 pandemic. In August 2021, Yassin had to step down, after 17 months in power, succeeded by Ismail Sabri Yaakob.

These subsequent collapses can be attributed to the historical polarization of Malaysian society, as Malaysia is divided along religious and ethnic lines – its east/west divide in particular – and, also, due to differing approaches to politics (Welsh 2020). The Malay majority, which composes around 50 % of the overall population, enjoys better access to power, law protection, and societal status, while Chinese and Indian Malaysians are subject to discrimination (ibid.). This political fragmentation is enhanced by the role of social media, where different groupings and elites deliver alternative messages.

The country has major newspapers in various languages: Malay, English, Chinese, and Tamil. Historically, freedom of the press has been limited, with numerous restrictions on publishing rights and information dissemination. Malaysia currently ranks at place 119 in the 2021 Press Freedom Index. This ranking has

fluctuated, as it saw an improvement in 2020, reflecting a relatively moderate degree of press freedom compared to its neighboring countries. However, in the following year a more authoritarian rule was re-introduced in the country, falling 18 places in the index, due to the policies of the Perikatan Nasional government. The main newspapers are owned by the government and political parties in the ruling coalition, although some major opposition parties also have their own, which are openly sold alongside regular newspapers.

45. Malaysia's most popular news/editorial websites (left) and online forums / social websites (right) according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



There is an east–west divide in the Malaysian media landscape, as Peninsular Malaysia accounts for the around 82 % of Malaysia's population and economy, and the Peninsular media gives low priority to news from the eastern side of the country, often treating the eastern states of Borneo as adjunct colonies to the Peninsula (Fernandez 2010). There are also tensions between Indonesia and Malaysia, that are sometimes said to be inflamed by media discourse.

Malaysian access to the internet has been gradually liberalized after the state's monopoly began to be removed after 2010, and like many of its neighboring countries in the South-East Asian region, mobile internet is very popular. 75–85 % of traffic to the popular news sites in our overview comes from mobile devices.

Mexico

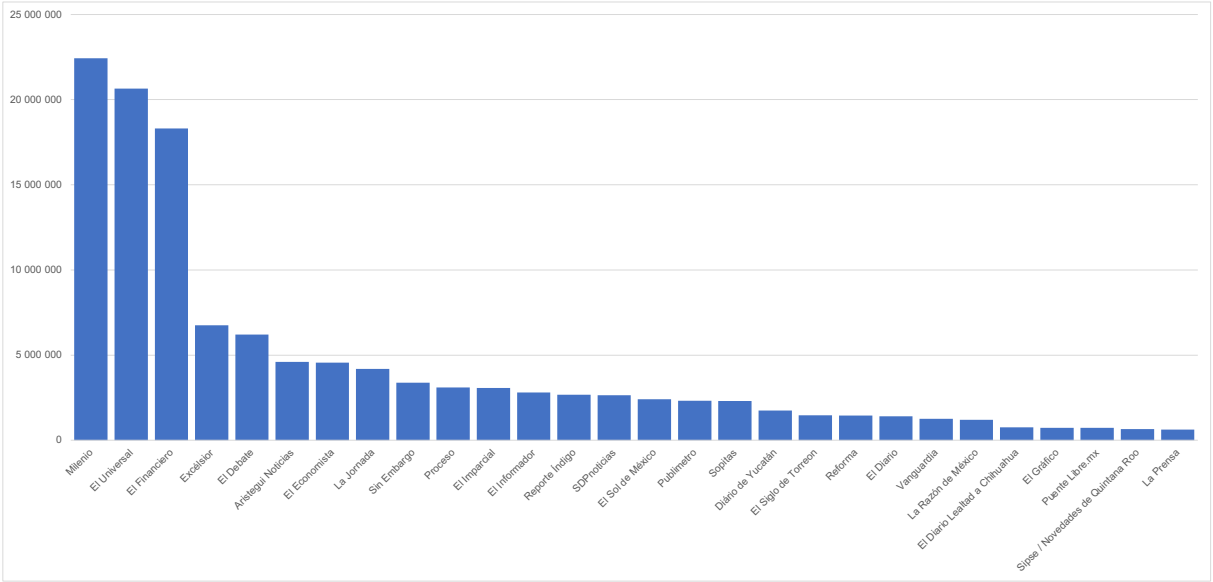
Population: approx. 130 million (2021)

Comparative digital news media penetration coefficient (our estimate): **0.72**

The past ten years have been marked by significant degrees of violence in the country, chiefly as a result of the influence of drug cartels and corruption into daily life, especially in some regions that are worse off than others. The political system of Mexico is characterized by considerable degrees of presidentialism (see Graph 27). During 2006–2012, Felipe Calderón was head of state, followed by Enrique Peña Nieto (2012–2018), and the current incumbent Andrés Manuel López Obrador. Calderón's presidency was marked by criminal justice reforms and a declaration of war against crime cartels (Economist 2016). Enrique Peña Nieto, as the head of the Institutional Revolutionary Party, started his term in office with promising reforms, which, however, were sidelined by corruption scandals and human rights violations. The widespread corruption, violence, rule-of-law deficits, and human rights abuses present significant societal challenges (Freedom House 2017), despite many anti-drugs and anti-cartel campaigns.

During the period under scrutiny, the indexes of liberal democracy and presidentialism have stayed relatively stable (Graph 27), as has the country's status as an electoral democracy. Notably, the power distribution seems to have become more balanced, while power appears to have shifted from the majority's monopoly to an "almost equal" society.

46. Mexico's most popular news/editorial websites according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



In terms of the media landscape, the domestic market for online news is dominated by three titles: Milenio, El Universal, and the business oriented El Financiero. Alongside these three stalwarts, there is a long range of medium-sized titles, indicating a relatively strong online media sphere, but still in an altogether different league than European or Northern American media markets. Media television duopoly Televisa / TV Azteca controls 99 % of the market (Huerta-Wong & Gómez 2013). Investigative journalists (Tuckman 2012) have shown that Televisa has been systematically biased in favor of Enrique Peña Nieto of the PRI party and acted to delegitimize his left-wing opponent Manuel López Obrador.

Mexico's overall digital news media penetration rates are comparable to those of Brazil, Indonesia, Malaysia, and Egypt. Some of the papers that have massive paper editions (photo-based sensationalist tabloids like La Prensa, for example) seem to have relatively small online circulations, by comparison. For the popular sites, around 75–85 % of traffic comes from mobile devices. For some websites, the share of mobile traffic is lower (e.g., El Grafico, at around 60 %, and Reforma, at around 54 %). Also in the online media sphere, there have been severe doubts about the legitimacy and fair play of Mexican politicians during the 2012 elections. Treré (2018) notes that the intensified use of digital technologies hardly corresponded with an increase in democratic participation, but rather that a set of (nowadays familiar) tactics were employed: fake online followers, fake interactions, digital bots, etc., accompanied by even more severe attempts like orchestrated attacking and blocking of oppositional activists in social media.


Nigeria

Population: approx. 206.2 million (2021)

Comparative digital news media penetration coefficient (our estimate): **0.12**

From 2010 to 2015, Goodluck Jonathan served as Nigeria's president. He was the first incumbent democratically elected president to be defeated in relatively free and fair elections in the country (Akinola 2016). The Nigerian landscape of the last ten years has been characterized by, among many things, the Boko Haram uprising – an Islamist movement in the north-eastern part of the country, perpetrating violent attacks on civilians and military. Some of the factors explaining Jonathan's electoral loss can be attributed to a high level of corruption, the escalation of terrorism, and mass kidnappings (Biakolo 2021).

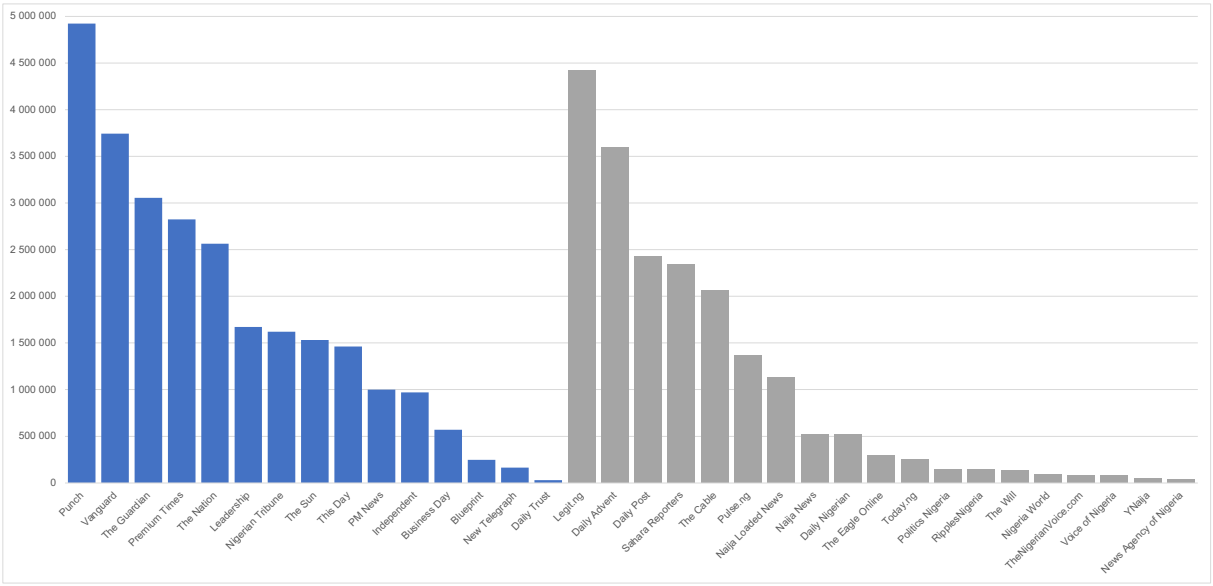
Since 2015, elections have become more competitive and the regime more liberal. 2015 was the first time in the country's history when an opposition candidate – Muhammadu Buhari (the current president) – won the presidential elections. As a former general, he campaigned for security provision, especially in the country's conflict zone in the north-east (Biakolo 2021). While northern Nigeria is largely Muslim, southern Nigeria is largely Christian, there is no outright conflict between these two regions, even though there are evident different political and religious orientations and ideologies. Buhari has not managed to reduce the economy's dependency on oil, nor expand the economy or suppress corruption (ibid.). From 2015 onwards, the government has also begun filtering domestic internet access to an increasing degree; our data also indicates an increase in presidentialism (Graph 27). At the same time, use of social media to organize political actions has become more commonplace, resulting in some noteworthy contemporary developments where political conflicts are acted out in social media. For example, in the recent year the government has begun striking down on social protest, to such a degree that militarized police are shooting protesters in the street – notably, the Special Anti-Robbery Squad (SARS) lies behind a lot of such extrajudicial abuses. Citizens have begun using social media to document the numerous violent crimes perpetrated by SARS officers. Activists rally around the so-called End SARS movement (a slogan coined in 2017 as a Twitter campaign, using the hashtag #EndSARS), which Buhari's government has tried banning, for example by shutting down Twitter in Nigeria.

Nigeria ranks very low on the World Press Freedom Index; it ranks at 120 out of 180 countries in the world rankings in 2021, and a lot of the problems stem from, the authoritarian military/police/politics complex and corruption – often a combination of both.

Being the most populous country in Africa, Nigeria is known for its superabundance of various spoken languages and dialects, with over 500 spoken native lan-

guages. Due to the strong historical links to the colonial legacy of the former British Empire, English nevertheless remains the official language. Due to the global reach of English, it remains the key language in business discourse and the very most prominent language for news and online text. The remaining colonial links are also seen in the news ecosystem, as some of the rare online news services offered in domestic languages like Hausa, Yorùbá, and Ìgbò are from the BBC World Service.

47. Nigeria's most popular news/editorial websites (left) and editorial aggregators / web portals (right) according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



One observation that we make, looking at the accessible Web traffic metrics, is that the list of small-scale news-aggregation websites is very extensive in Nigeria, compared to other African nations. English being the main written language in Nigeria, means that an undergrowth of such sites (many of them most likely highly automated) aligns with the global circulation of English-language, machine-readable web content.

Nevertheless, compared to other African nations, the online news media in Nigeria command astonishingly small audiences, when put into relation with the country's large population. The level of literacy and access to the internet will be a contributing factor to Nigerians' small audience to online media platforms, compared to size of population. Partially it might also be an artefact of the specific ways in which SimilarWeb measure web traffic. According to SimilarWeb,

Kenya's largest online news outlet, Tuko, has just under two million unique visitors per month, while Nigeria's biggest online newspaper, Punch, has only around 2.5 times as many unique visitors per month — while Nigeria has almost four times the population. In Brazil, a country with a similarly large population as Nigeria's, the largest news/current affairs aggregators each command over 90 million unique visitors per month. In comparison, Nigeria's online news media therefore appear to be a cottage industry; there really are no large online news aggregators or newspapers of the same kind as in Brazil.

As much as 92–95 % of traffic comes from mobile devices. Notably, large portions of web traffic also come from other jurisdictions than the Nigerian mainland; significant traffic comes from the US and the UK, something which is primarily explained by the considerably large English-language Nigerian diasporas in these countries. For the news aggregator sites this was particularly notable. One example is the very popular news aggregation website Legit.ng, receiving a lot of its traffic from the US and Canada, while The Cable receives a lot of its traffic from Germany. Interestingly, Legit.ng also crops up in international measurements of popularity of Facebook Pages. According to analytics platform Newswhip's measurement of engagement rates of publishers on Facebook in January 2022 (Nicholson 2022), Legit.ng came in fourth place, after Dailywire.com, BBC.co.uk, Dailymail.co.uk, and CNN.com. According to this measurement, Legit's Facebook page saw over 14 million user engagements over the course of one month. This information indicates that online media use in Nigeria is, as in many countries, largely a matter of interacting with legacy media, but on social media platforms, a form of audience behavior not easily captured by raw web traffic data such as that we use in this report.

Poland

Population: 159pprox.. 37.8 million (2021)

Comparative digital news media penetration coefficient (our estimate): **1.31**

In the recent decade, Poland has seen a deterioration of the rule of law, especially in the last five years, since the right-wing national-conservative Law and Justice (PiS) party formed a majority government. The party hosts a clearly nationalist-chauvinist agenda, catering to conservative Catholics, pursuing Islamophobic, anti-LGBT, and anti-abortion rhetoric. Between 2005 and 2007, PiS held a brief coalition government, together with marginal Eurosceptic parties. Lech Kaczyński served as president and his brother, Jarosław Kaczyński, as prime minister. In the 2007 elections, the centre-right party Civic Platform won the 2007 elections, making Donald Tusk the prime minister and seeing a succession of more Euro-friendly cabinets (Casal Bértoa 2021).

PiS won the 2015 presidential and parliamentary elections, precipitating a turn to the right in Poland, as candidate Andrzej Duda won over the incumbent, Bronislaw Komorowski, in the presidential election and PiS candidate Beata Szydło became prime minister in 2015–2017, succeeded by Mateusz Morawiecki.

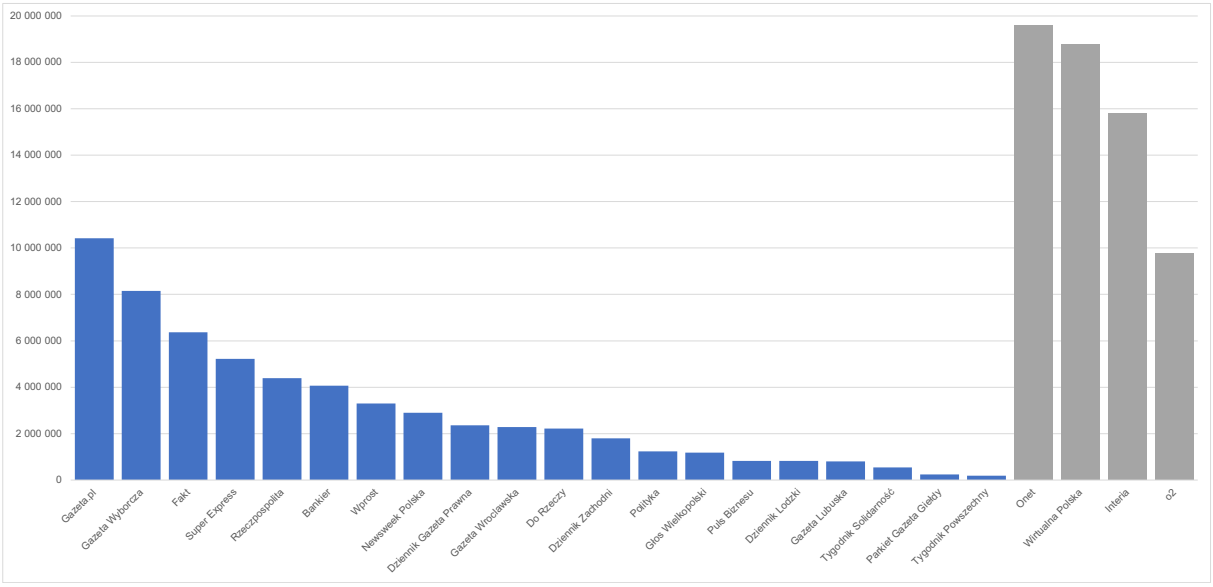
Since these electoral victories of the PiS, a decrease of liberal democracy can be seen in the index scores. Index values for Freedom of Expression and Alternative Sources of Information have considerably dropped since 2015, and Polish society appears to have become more polarized on major political subjects (Graph 25). New laws were passed to limit the power of the Constitutional Court and enabling the government to appoint managers of public-service broadcasting. Since 2016, there have been waves of protests against the government's attempts to curtail abortion rights.

Despite these illiberal trends, political corruption does not seem to have worsened in Poland as rapidly as in Hungary, and the noted index values for corruption are lower than in Italy. Nevertheless, according to Transparency International, Poland's administration has become significantly more corrupt in the last years, dropping from a score of 63 % on the Corruption Perceptions Index in 2015 to 56 % in 2020. Transparency International reports that PiS has consistently promoted reforms that weaken judicial independence, and that the Covid-19 crisis was used as an excuse to amend and repeal hundreds of laws, limiting access to information for citizens and journalists, and allowed for new arrangements of opaque public spending.

Historically, Poland is a country with a very strong clerical/legal/administrative legacy; it is one of the early European countries to have a culture of letters and newspapers. Already in the 17th century there was a significant domestic print

culture, with titles like Merkuriusz Polski Ordynaryjny. This is still visible in that many of the more serious news media in Poland seem to have a clerical and legal orientation, in addition to the internationally familiar business news format.

48. Poland's most popular news/editorial websites (left) and editorial aggregators / web portals (right) according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



The Polish media market is the largest domestic media market in Eastern Europe, with almost 38 million Polish-speaking inhabitants, and millions more in the neighboring region – and a relatively homogenous national culture, highly influenced by Catholicism. There is a strong legacy of local/regional press; all Polish regions have their own newspapers, mostly limited to the province where they are issued. In addition, the major national newspapers often issue daily attachments related to local topics. Across the Polish media market, there is quite significant diversity of editorial stances, ranging from far-left to far-right; nevertheless, since the current national government is very rightwing, there is a rightward trend also in the media landscape.

There is grave concern, both domestically and in the international community, regarding recent political developments in Poland. In February 2021, many of the private media companies in Poland collectively protested government plans to impose a tax on advertising revenues, which many of the media corporations saw as a targeted attack on independent journalism. Also, the market-leading internet portal Onet took part in this protest.

> Broadcasters replaced TV shows with black screens with the message "This is where your favourite programme was supposed to be"; internet portals blocked access to articles; and 43 media groups signed an open letter branding the plan "extortion". They warned that its introduction would lead to the "weakening, or even liquidation" of some Polish media companies, whose budgets have already been shredded by the coronavirus pandemic. (Shotter 2021)

Since PiS took office in 2015, Poland has fallen from the 18th to the 64th place in the World Press Freedom Index. A very significant drop, and the country's press freedom now ranks below countries like Niger and Armenia. While Poland has had a strong domestic media market for decades, the state-affiliated public service broadcaster (TVP) has, since PiS entered government, become a propaganda tool of the state, and many observers now fear that also private media companies are being entered into the fold of government influence. PiS politicians seem to have a policy of wanting to make state-controlled companies buy up foreign-owned Polish media; in December 2020, for example, state-owned oil refiner PKN Orlen acquired Polska Press – controlling over 20 of Poland's 24 regional newspapers, and almost 120 local weeklies – from Germany's Verlagsgruppe Passau (Shotter 2021). This falls into the pattern, endemic in Russia and Belarus and now increasingly common also in, e.g., Hungary and Poland, of directly state-affiliated or state-owned corporations seizing ownership of popular editorial media brands, a tendency "more akin to Gazprom's purchase of Russian media groups for Vladimir Putin, than to Bezos's purchase of the Washington Post" (ibid.).

Importantly, many media companies are indirectly supported by government through the substantial advertising spend of state enterprises and agencies. Polish researchers have shown that the main beneficiaries are pro-government titles such as daily newspaper Gazeta Polska Codziennie, and weekly magazines Sieci and Do Rzeczy. For papers like these, state-related revenues accounted for significant shares of total ad revenues (Wiśniewska 2019).

In the indexes, it can also be gleaned that online media has fractionalized more after 2017. This is also a point in time after which presidential power has increased. From 2015 onwards, a significant rise in the use of social media for protesting is noted. For the popular websites, around 57–77 % of traffic comes from mobile devices. For some websites the share of mobile traffic is lower (e.g., Rzeczpospolita, at around 54 %). For the major internet portals (Onet, Wirtualna Polska, Interia) traffic is notably split at around 50 % mobile, 50 % desktop.

In our traffic data we see a similar phenomenon as in some of the other 19 countries, where the very largest websites providing news and editorial text are the online portals, acting as news aggregators. Market leader Onet is one such portal, established in 1996 by famous Polish video game company Optimus (nowadays known as CD Projekt), acquired by German-Swiss media conglomerate

Ringier Axel Springer in 2012. Onet also used to host individual blogs (reportedly, up towards 800,000 active blogs) – but in 2018, due to the shifts in the social media outlined in this report, this service was closed. Wirtualna Polska is also a web portal, founded in the early days of the Web (1995), being the first webpage to gain mass popularity in Poland. Wirtualna Polska offers a range of services, including news, e-commerce, and advertising. Interia, the third largest web portal, was established in 2000 as part of a joint venture of the leader of the Polish IT market, Comarch SA, and the largest Polish radio station, RMF FM. The website offers email service, web hosting, and domain name registration, online games, blogs, chat rooms, internet forums and streaming media. Owned by German multimedia conglomerate Bauer Media since 2008.

The biggest online newspapers are Gazeta Wyborcza and Fakt, both centrist/liberal in their political leanings, the former a more serious broadsheet and the latter more sensationalist/populist in its reporting. Gazeta.pl is a web portal, which is however an offshoot of Gazeta Wyborcza and thus counted as a news provider in the blue section of our graph. The Wyborcza.pl URL operates with a paywall model and contains articles from the paper edition as well as current news and comments on current events. Gazeta.pl and Wyborcza.pl have been cooperating with each other for years. Gazeta.pl focuses on free, internet-friendly content with mass appeal, while Wyborcza.pl focuses on premium content.

Russia

Population: approx. 146 million (2021)

Comparative digital news media penetration coefficient (our estimate): **1.48**

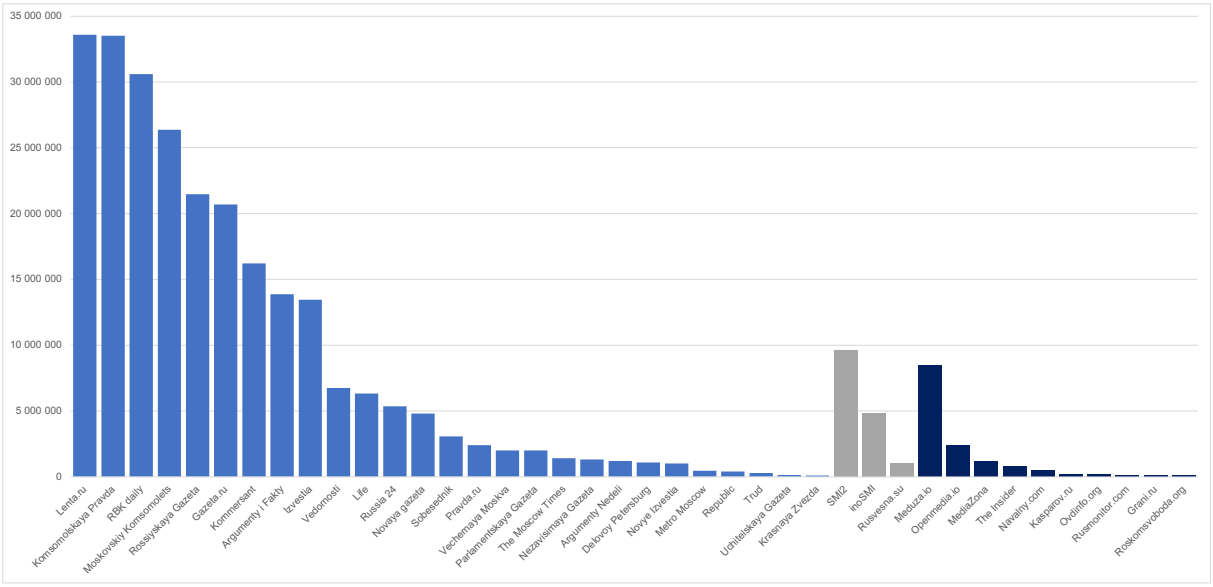
NB: As this report is published, we are in the midst of a Russian invasion of peaceful neighboring country Ukraine, a strategic move instigated by Russian tyrant Vladimir Putin with unclear world-historical repercussions. Since the turn of the millennium, Russian politics have been characterized by Putin's presidency, as he has continually ruled the country since the shift of the millennium, only stepping out of office during 2008–2012 to briefly let his ally Dmitry Medvedev serve as president. Under Putin, the trend in Russia has been clear: increased centralization and authorization (Freedom House 2021). Since 2012 there has been a crackdown on freedom of expression, with numerous laws passed that act to limit democratic freedom of speech and freedom of assembly. Thus, "state intrusion in media affairs has reached a level not seen in Russia since the fall of the Soviet Union" (Human Rights Watch 2017). In the academic community, Russia is therefore clearly no longer considered to be a democracy; the regime uses strategies like imprisonment of political opponents, purges, repression, and prohibition of a free press, as well as a lack of free and fair elections. In the corruption and press freedom indexes, Russia is trailing far behind European countries; it ranks as only having a 30 % score on the Corruption Perceptions Index (place 129 in the world), meaning that corruption in Russia is comparable to countries like Mali and Azerbaijan, while nearby countries like Poland and the Baltic states are significantly less corrupt.

The 2010s saw a continuation of the Putin regime's corruption and political violence, even murder. In the 2011 parliamentary elections, Putin's pro-government party United Russia captured fewer votes than previously. Mass protests were held against electoral fraud. In 2012, Putin became the president for the third time, after Medvedev's intermission. Simultaneously, a new law was introduced, classifying NGOs receiving money from abroad as foreign agents. The year 2014 became a crucial juncture in Russian history because of the outbreak of war in Ukraine and the annexation of Crimea. In 2018 Putin became president for the fourth time, initiating a constitutional referendum allowing him to stay in power until at least 2024. In 2020, the opposition politician Alexei Navalny was poisoned by pro-government thugs and sentenced to prison months later.

Putin's model of government has, by his supporters, been seen as a paradigmatic alternative to Western deliberative democracy. At the same time, scholars have characterized his rule as one of "state capitalism," similar to China's maintenance of financial stability through authoritarian control of democracy and corporations and oligarchs kept in tight collaboration with government agents.

Considering the non-democratic trend, it is striking that among our 20 selected countries, Russia is the country with the highest average use of social media to organize political action on record – an indicator that has been continuously rising since 2010. At the same time, Russia, along with Sweden, appears to be the *least* polarized societies on major political issues in our selection. This reasons behind this metric are probably somewhat different in Russia compared to Sweden. If Sweden is a country with very high rates of civic trust, and trust in government, alongside a political culture of relative consensus, Russia's vast text-based culture is often said to allow for certain degrees of dissent. However, the Russian political system is also characterized by the highest score of presidentialism and corruption among all our selected countries, indicating that repression and policing of dissent is critical in the country, meaning that the *de facto* degree of political dissent is restrained.

49. Russia's most popular news/editorial websites (blue), editorial aggregators / web portals (gray), and explicitly oppositional news / editorial websites (dark blue) according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



There is a very strong print culture in Russia, with over 400 daily newspapers and, reportedly, the largest number of newspaper journalists in the world (approx. 102,000), followed by China (approx. 83,000) and the United States (approx. 54,000), according to UNESCO statistics in 2005 (Treisman 2011: 358). Nevertheless, press freedom and freedom of speech in Russia is very bad. The country now ranks at the bottom of the World Press Freedom Index; place 150. Along the overextended period of Putin's presidential tenure, the country's ranking has continually fallen in international indexes of press freedom. Over five years, between

2011 and 2016, the government forced changes of ownership over several significant newsrooms with pan-Russian reach, all of them previously associated with independent reporting. RBC, Forbes, Russian Media Group, TV2, Russkaya Planeta, REN TV, Grani.ru, Lenta.ru, TV Rain (Dozhd), RIA Novosti, Gazeta.ru and Kommersant were suppressed or taken over. Censorship was achieved through different techniques – sometimes through government taking control of the company shares, passing ownership and management to newly created institutional bodies, controlled by state-approved managers (e.g., RIA Novosti), while TV Rain was forcibly removed from TV channels and only allowed to continue business as an Internet-only station. All but one national TV channel are fully or partially owned by the state. The remaining channel, NTV, is owned by Gazprom, in which the state has a controlling stake. The situation in the radio market is similar.

State censorship of the online news domain was supercharged during the 2014 Russian military intervention in Ukraine and the occupation of Crimea. On March 13, 2014, access to publications like Daily Journal (ej.ru), Grani.ru, Kasparov.ru and Alexei Navalny's LiveJournal blog was blocked by government agency Roskomnadzor at the request of the General Prosecutor's Office of Russia. In 2020, the ECHR recognized that the blocking of Grani.ru was contrary to the European Convention for the Protection of Human Rights and Fundamental Freedoms and ordered the Russian government to pay a fine of €10,000 to the publication.

Communications regulator Roskomnadzor seems to have increased its hold even further, in recent years. In December 2021, Russia's supreme court ordered the closure of the country's oldest human rights organization, Memorial International (Roth 2021). The same month, Roskomnadzor blocked the website of OVD-Info after a court ruling. For years, OVD-Info has been an important civil society organization, documenting anti-Kremlin protests and providing legal support to victims of political persecution. During 2021, Roskomnadzor has also been targeting international social media companies. The regulator was, for example, deliberately slowing down Russian internet users' access to Twitter, as the authority alleged that immoral content circulates on the social media site (e.g., child pornography, drug- and suicide-related information; Stolyarov 2021). This increased internet censorship is most likely a reaction to the civic protests demanding the release of the jailed opposition leader Alexei Navalny. Since 2019, there is even a series of legal amendments (the so-called "sovereign Internet" law) that, in theory, enables Russian authorities to isolate the country's internet backbone (see, e.g., Human Rights Watch 2017). This situation, forcing large independent media companies to run their operations from abroad, was notable to us as we tallied the different media titles, and this is the reason why we have sorted our data in two bins: 'state-sanctioned' versus 'oppositional' titles.

For the popular Russian news sites, around 50–70 % of traffic comes from mobile devices. News aggregator SMI2 has a lot of mobile traffic (78 %) while some other sites have a notably smaller share of traffic from mobile, compared to desktop. Some websites have less than 50 % of their traffic coming from mobile, compared to desktop – notably, large aggregators like inoSMI and Rusvesna.su seem to have predominantly desktop-based traffic. Russian news sites often have sizeable audiences also in the neighboring Russian-speaking countries. Some of the news websites seem to be more transnational than others in their reach; Svoboda.org, for example, seems to have significant traffic from other countries than Russia.

In our quantification of traffic data, we have tried to compare the state-sanctioned, mainstream media outlets with those that have an explicitly oppositional profile. The largest news media titles to emerge are Lenta.ru, Komsomolskaya Pravda, RBK daily, Moskovskiy Komsomolets, and the government's own Rossiyskaya Gazeta – all very familiar news brands in Russia, with notably short and simple web addresses. The biggest news aggregator is SMI2, and the biggest oppositional news site is Meduza.io.

Lenta is an interesting case, compared to many of the very established legacy newspapers. An online-only commercial newspaper, it was founded in 1999 and was actually condemned and ostracized by the government in 2014, due to its coverage of the conflict in Ukraine. However, subsequently the newspaper saw a major reshuffle of staff and management, in order to take a more Kremlin-friendly editorial approach; many of the defectors who resigned founded the alternative online title Meduza, which is critical of the current Russian government, and operates out of Riga, Latvia. Up until 2021, Meduza was only an aggregator, but as of recently it also publishes original content.

South Africa

Population: 167pprox.. 60.3 million (2021)

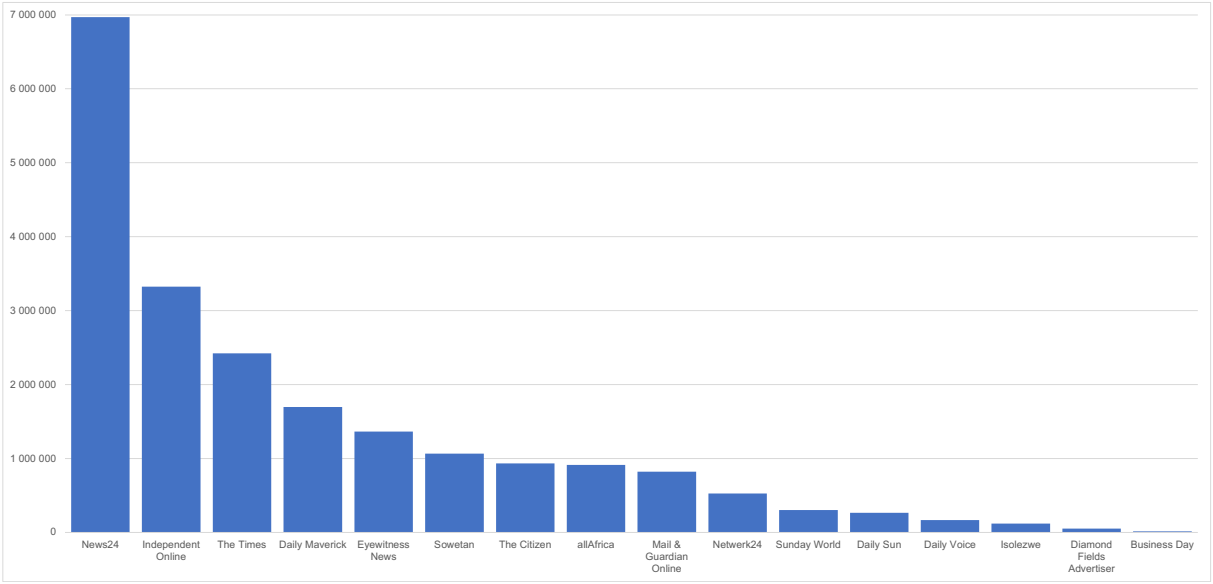
Comparative digital news media penetration coefficient (our estimate): **0.33**

Since 2010, the mode of governance in South Africa has been transformed from liberal democracy to merely an electoral one. In 2009 Jacob Zuma was elected as a president and ran the country until 2018, when he resigned after several corruption scandals. In the 2010s, the country witnessed many protests, due to factors like the shrinking economy, inequality, poor service delivery, and other social reasons.

After Zuma's resignation, parliament elected Cyril Ramaphosa as president in 2018. The incumbent has inherited a divided party system, a shrinking economy, and widespread corruption. However, the freedom of expression and alternative sources of information remain pretty stable in the county, and its values on the index are relatively high, comparable both with other African states and the US.

The South African party system is, at the national level, characterized by one-party dominance – the African National Congress (ANC). The electorate continues to vote largely along racial lines, where ANC claims to represent the black majority (Brooks 2004). In recent years, ANC's political dominance has been challenged, primarily at provincial and municipal level, by new parties such as the moderate Democratic Alliance (DA) and the radical Economic Freedom Fighters (EFF).

50. South Africa's most popular news/editorial websites according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



In the pan-African context, South Africa is a hub for media companies – arguably more so than other dominant regions, such as Nigeria. For this reason, some of the metrics in our overview might divulge Web traffic from numerous countries across the sub-Saharan African region, not only domestic South African audiences. However, the issue with transnational media reach is perhaps more pronounced with broadcasting media (radio, tv) which are not part of our investigation. The popular news website AllAfrica, for example, is an online news provider with explicit pan-African ambition and transnational reach. The company has offices also in other African countries.

Media24 is the most dominant media owner, in turn a part of Naspers, a multinational group of multimedia and e-commerce platforms. Among the titles owned by Media24 are News24 (as we see in our graph, South Africa's leading news website by a wide margin) and Daily Sun. Many of its other media titles (Volksblad, The Witness, Beeld, Die Burger, Rapport) are legacy newspapers that have traditionally had paper editions but are increasingly digital-only, and those are collectively represented by the Netwerk24 webpage, also owned by Media24. A similar arrangement is found in the second-most popular news site, Independent Online, which hosts online versions of a number of South African legacy newspapers (e.g., The Star, Pretoria News, Cape Times, The Daily Voice, etc.).

For the popular sites, around 66–80 % of traffic comes from mobile devices.

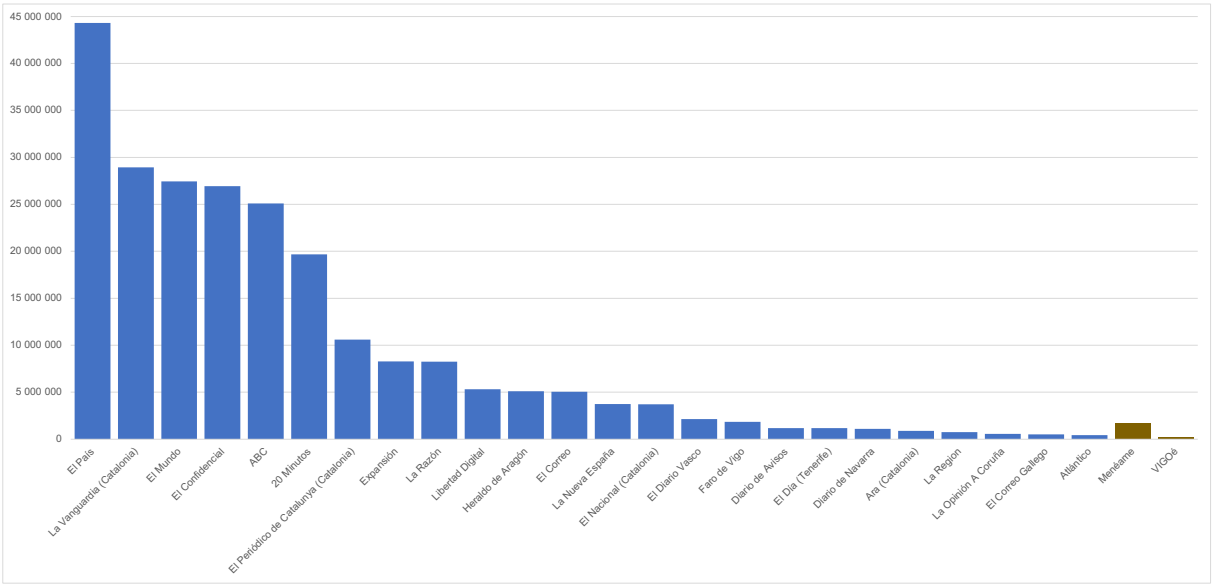
Spain

Population: approx. 46.8 million (2021)

Comparative digital news media penetration coefficient (our estimate): **4.38**

During 2010–2020, there has been four cabinets in Spain, run by two prime ministers: Mariano Rajoy Brey from the People's Party and Pedro Sánchez Pérez-Castejón from the Socialist Worker's Party (Casal Bértoa 2021). During this decade, political polarization in Spain has increased considerably (Graph 25) while governance indexes show relatively constant trends. The financial crisis brought up harsh economic consequences and caused waves of protests in 2010. In 2013, unemployment in the country reached a peak of over 27 %. The rise of populism can be seen in 2015 with the appearance of popular anti-austerity movement Podemos on the political landscape. The Spanish political scene changed with the 2017–2018 constitutional crisis, caused by the two failed Catalonian independence referendums, and the ensuing protests. Since then, Spanish nationalism has grown, evidenced by the entrance of far-right party Vox entering the Congress in 2019. At the same time, the population feels great distrust against the media; according to the Eurobarometer on Media pluralism and democracy (European Commission, 2016), Spain ranked 27th among the 29 EU countries in terms of citizens' perception of the diversity of views and opinions in the media (Salaverría & Baceiredo n.d.).

51. Spain's most popular news/editorial websites (left) and online forums / social websites (right) according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



In media-related indexes, Spain not only has notably strong penetration rates of its domestic online news/editorial websites compared to population, it also ranks highest among our 20 countries regarding the use of social media to organize political action. Interestingly, in our report, this can be related to the observation we make (see Graph 32), that Spain has a significant online news presence, relative to population. One of the popular movements in Spain during the time period we are interested in (2010–2020), Indignados, emerging in 2011, alongside the regionally popular separatist movement for Catalan independence, can be seen as protest movements that have left a considerable mark on online political discourse during this timeframe. Spain also appears to have one of the most dramatic increases in polarization on major political issues since 2012. From this time point, the Spanish state is also found to have increased its attempts in filtering the Internet.

Following the death of dictator Francisco Franco in 1975, the Spanish media landscape underwent a dramatic transformation as press freedom and freedom of speech allowed for new papers and a surge of audiovisual corporations. Today, the media landscape offers outlets with a range of ideological positions, and some of them reflect the polarized and divided political climate of a country that consists of 17 autonomous communities. Mainly, news media across Spain could be seen as being divided between constitutionalist media (sometimes monarchist, and in support of national unity) and nationalist media (pro-autonomous in for instance the Basque Country, or Catalonia) (Salaverría & Baceiredo n.d.).

The most read papers in Spain are usually said to be the three Madrid dailies: El País, El Mundo and ABC, which are controlled by PRISA, Unidad Editorial and Vocento. Out of them, ABC the largest legacy paper that operated throughout the Civil War (1936–39), and during the Franco Regime. It is also the most conservative of the three, with close ties to the People's Party. Today, the three papers are challenged by free newspapers such as 20 Minutos and online papers such as El Confidencial, and a number of smaller publications that have been promoted by their own off-shoot journalists, coming from established papers.

For the popular websites, around 73–85 % of traffic comes from mobile devices. Quite a lot of traffic comes from Spanish-speaking countries overseas: Mexico, Argentina, Colombia, the US. Especially the leading titles – El País and La Vanguardia – appear to have significant audiences abroad.

Regarding domestic social media platforms, Menéame is worth a mention. It uses a similar format as Digg and Reddit, with community participation in the form of web links and discussion around them.

Sweden

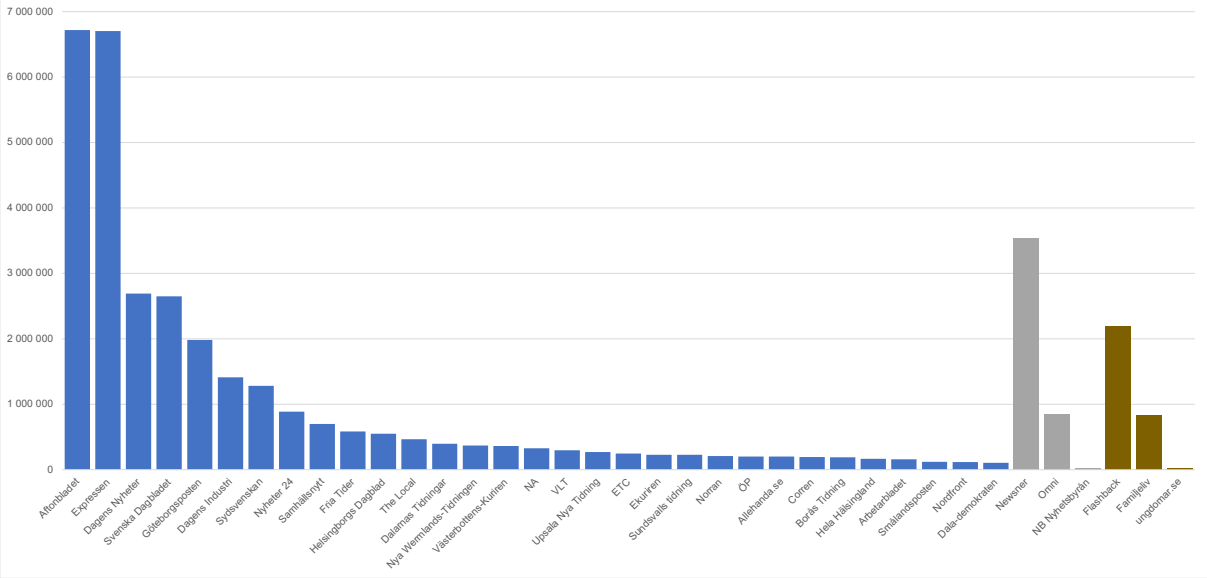
Population: approx. 10.1 million (2021)

Comparative digital news media penetration coefficient (our estimate): **2.53**

During the period under observation, political developments in Sweden have not been very dramatic. The center-right government of Fredrik Reinfeldt was replaced by a Social-Democrat led one in 2014. The only indicators that have changed somewhat in Sweden for the period under scrutiny are the polarization of society on major political issues, and the domestic distribution of political power: Sweden has become slightly more polarized since 2014, but compared to other countries, polarization is still very low. Meanwhile, a shift in power distribution occurred, already at around 2012, from "equal" to "almost equal" (Graph 30). Immigration to Sweden has been continuously high in recent years, reaching a peak of 163,000 people applying for asylum in Sweden in one year (2015) while, it should be added, only 33,000 were granted asylum that year.

Interestingly, Sweden and Kenya are the only countries, out of our 20 selected countries, where the average use of social media to organize political action has not increased during 2010–2020. In general, the indicator values for this parameter are relatively low, indicating relative political passivity among the general online population. Sweden is also characterized by one of the least polarized party systems in Europe, meaning that the ideological distance between parties is relatively small (Casal Bértoa 2021). During the studied period, three different cabinets have been in charge of the country. Fredrik Reinfeldt from the center-right Moderaterna party acted as prime minister from 2010 to 2014 and was succeeded by social-democrat Stefan Löfven (the current prime minister).

52. Sweden's most popular news/editorial websites (blue), editorial aggregators / web portals (gray), and online forums / social websites (brown) according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



Notably, populist-right media outlets are visible in our data table for most popular online news outlets. This is something of a shift in Sweden, in recent years, at least in that the key titles among the populist-rightwing press, sometimes referred to as 'alt-right' media, have become legitimized as formal parts of the press system, due to a process of increased media formalization, requiring titles to have official directorship of publication in order to receive state subsidies, as the system for state subsidies was revised in 2016 and 2019.

USA

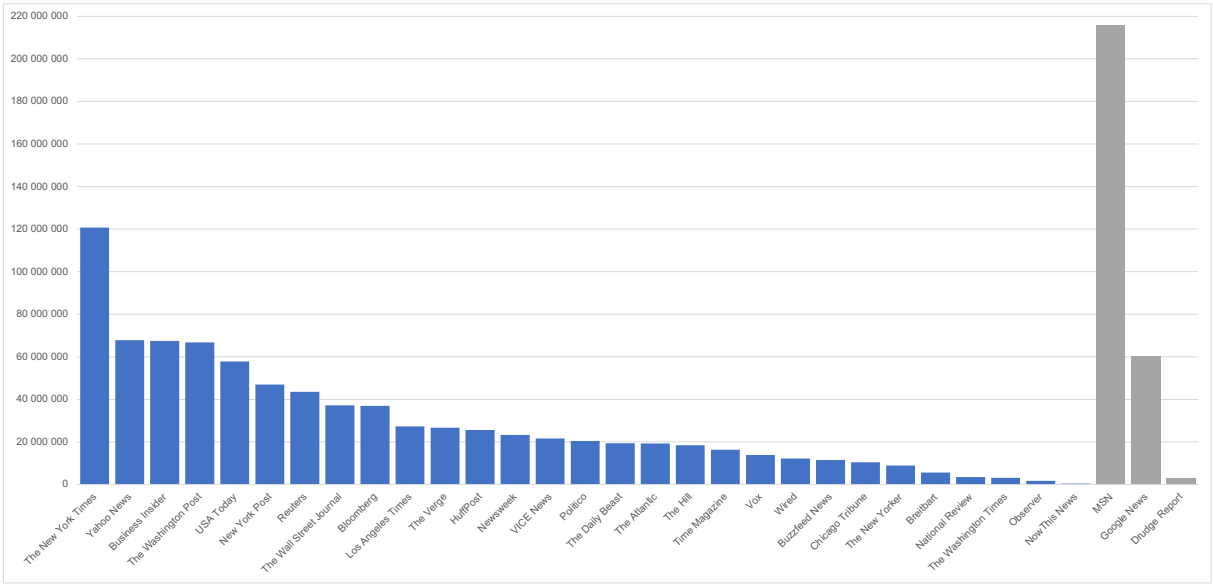
Population: approx. 332.8 million (2021)

Comparative digital news media penetration coefficient (our estimate): **1.53**

Since the 2016 elections, ushering in the one-term presidency of Donald Trump, American society has become characterized by greater polarization on major political issues (Graphs 8, 21, 25), albeit macro values do not indicate as extreme a polarization as in Brazil, Spain, Hungary, and Poland. Since 2016–2017, a slight shift on the presidential scale (Graph 27) is also spotted, indicating an increase in presidentialism. In recent years, popular resistance movements such as Black Lives Matter have emerged, alongside protests following the 2020 elections, culminating in the storming of the Capitol in January 2021. Donald Trump was succeeded by Democrat Joe Biden, who won the 2020 presidential elections, entering office in 2021.

In the most recent Reuters Institute Digital News Report (Newman et al. 2021), the USA is seeing the lowest score of general trust in the media of all 46 countries surveyed. It should be noted, however, that Americans' trust in the mass media has rebounded slightly after hitting a record low in 2016 (Brenan 2020). The American news market remains highly polarized, Newman et al. note (2021: 113). Cable news channels Fox News, CNN, and MSNBC – together with some of the purely online news brands like Yahoo and BuzzFeed – have some of the highest levels of distrust. In the media field, countless observations could be made, but we will try to be brief.

53. USA's most popular news/editorial websites (left) and editorial aggregators / web portals (right) according to average monthly unique visits (SimilarWeb traffic data, July and August 2021)



Since the World Wide Web was launched in 1995, newspaper circulation in the US has more than halved. In 1995, 61 million daily papers were circulated, while 25 years later, the estimates were slightly above 24 million (Pew Research Center 2021). Meanwhile, the number of unique visitors to the 50 largest newspaper websites has risen quickly. Since the final quarter of 2014 the number has risen from 8.2 million to 13.9 million by the end of 2020 (ibid.). Meanwhile, internet media seem to be instrumental also for social mobilization; the usage of social media to organize political action has seen a clear upwards trend since 2014.

During the Trump presidency, the online news sphere saw further division than before, between liberal-leaning and conservative sites. Many papers and news websites saw a rise in readership and profits from the reporting of the White House and Trump's activities (Cocco 2018), and this period also saw the rise of new titles such as The Hill. Explanatory, social-media friendly news such as Vox took more space on digital platforms, often catering to younger progressives, sometimes called 'Generation Z' audiences. Legacy newspapers like The New York Times have fallen under growing critique from the left for reflecting the opinions the liberal 'coastal elites' (Roberts 2018, Frost 2019). Similarly, most American mainstream news outlets were during this time antagonized with the trope of "fake news," upheld by Trump supporters.

So-called 'woke' rhetoric on the left has clashed with radicalized political attitudes on the right, especially salient during the Black Lives Matter protests in 2020, debates around Covid-19 vaccinations, and possibly culminating with the alt-right insurgency leading to the January 6th storming of the Capitol.

Regarding our selection of sources for this report, similar arguments as those for, e.g., the UK can be noted; there is a measurement problem that is especially pertinent to English-language, Spanish-, Arab-, and Chinese-language websites, namely that some sites, while nominally based in one country, as regards their top-domain URLs, can be accessed, and might indeed have significant audiences, in entirely other countries. So, it is with this proviso one should approach the Web traffic metrics for the US-based sources.

In our graph, MSN comes out as the leading online news outlet, if we look only at the metric of recorded Web traffic. The New York Times also appears as a singular newspaper, in terms of reach. Nevertheless, there is a score of US-based online news providers enjoying a monthly average of 30 million unique visitors or more, some of these titles being aggregators (Google News, MSN) and the rest being producers of original news items (note that Yahoo News also features its own content). For the popular American news sites, around 57–67 % of traffic comes from mobile devices. There are some exceptions, however (e.g., MSN, New York Times, ABC News, Yahoo News, Google News, Wired, Wall Street Journal, New Yorker, Drudge Report, OAN) that all have a ratio of around 50 % mobile or less,

signifying an older and, in some of the cases, also rather wealthy demographic (affording laptop computers in addition to mobile phones). Also, Reddit sees a significant share of its traffic coming from desktop devices (66 %), rather than mobile (33 %).

US business and tech newspapers seem have rather transnational audiences (Bloomberg, The Verge, e.g., have considerably large overseas readerships) while outlets specializing in political polemic (e.g., Breitbart) seem to have more domestic audiences.

References

Akinola, A. (2016). Factors in the defeat of Goodluck Jonathan. *The Guardian* (Nigeria), September 16. https://guardian.ng/opinion/factors-in-the-defeat-of-goodluck-jonathan/ Accessed February 28, 2022.

Alba, D. (2021). Facebook sent flawed data to misinformation researchers. *New York Times*. September 10. https://www.nytimes.com/live/2020/2020-election-misinformation-distortions#facebook-sent-flawed-data-to-misinformation-researchers Accessed December 1, 2021.

Alexa (n.d.). What are unique visitors, pageviews, and visits? https://support.alexa.com/hc/en-us/articles/200462330-What-are-unique-visitors-pageviews-and-visits- Accessed December 1, 2021.

Al-Jazeera (2021). Russia blocks website of protest monitoring group OVD-Info. December 25. https://www.aljazeera.com/news/2021/12/25/russia-blocks-website-of-prominent-rights-monitor Accessed February 28, 2022.

Allen, J.; M. Mobius; D.M. Rothschild; D.J. Watts (2021). Research note: Examining potential bias in large-scale censored data. *Harvard Kennedy School (HKS) Misinformation Review*. DOI: .37016/mr-2020-74

Andreotta, M.; R. Nugroho; M.J. Hurlstone et al. (2019). Analyzing social media data: A mixed-methods framework combining computational and qualitative text analysis. *Behavior Research Methods*, 51:1766–1781. DOI 10.3758/s13428-019-01202-8

Ariely, G. (2015). Trusting the press and political trust: A conditional relationship. *Journal of Elections, Public Opinion and Parties*, 25(3): 351–367.

Associated Press (2013). Kenyan president accused of backing post-election violence that killed 1,000. *The Guardian*, 22 May. https://www.theguardian.com/world/2013/may/22/uhuru-kenyatta-election-violence-report Accessed February 28, 2022.

Bajomi-Lazar, P. (n.d.). Media Landscapes: Hungary. European Journalism Centre. https://medialandscapes.org/country/hungary. Accessed November 26, 2021.

BBC Media Action (2018). Kenya – Media Landscape Report. November. https://www.communityengagementhub.org/wp-content/uploads/sites/2/2019/09/Kenya-Media-Landscape-Report_BBC-Media-Action_November-2018v2.pdf Accessed February 28, 2022.

Biakolo, K. (2021). Nigeria's President Should Resign. *Foreign Policy*, April 29. https://foreignpolicy.com/2021/04/29/nigeria-buhari-should-resign/ Accessed May 3, 2021.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4): 243–257.

Bland, B. (2019). Politics in Indonesia: Resilient elections, defective democracy. Lowy Institute, April 10. https://www.lowyinstitute.org/publications/politics-indonesia-resilient-elections-defective-democracy Accessed February 28, 2022.
Bolin, G., & J. Andersson Schwarz (2015). Heuristics of the Algorithm: Big Data, User Interpretation and Institutional Translation. *Big Data & Society*, 2(2).
Boullier, D. (2017). Big Data Challenges for the Social Sciences and Market Research: From Society and Opinion to Replications. In: F. Cochoy, J. Hagberg, M. Petersson McIntyre & N. Sörum (Eds.) *Digitalizing Consumption: Tracing How Devices Shape Consumer Culture*. 20–40. Trans. J. O'Hagan. London, New York, NY: Routledge.
Brenan, M. (2020). Americans remain distrustful of mass media. Gallup, September 30. https://news.gallup.com/poll/321116/americans-remain-distrustful-mass-media.aspx Accessed February 28, 2022.
Briggs, A., & P. Burke (2020). *A Social History of the Media*. Fourth ed. Cambridge & Medford, MA: Polity Press.
Britannica (2021). Luiz Inácio Lula da Silva. Accessed May 3, 2021.
Bucher, T. (2018). *If…Then: Algorithmic Power and Politics*. Oxford: Oxford University Press.
Brooks, H. (2004). The dominant-party system: challenges for South Africa's second decade of democracy. *Journal of African Elections*, 3(2): 121–153.
Brownsell, J. (2013). Kenya: What went wrong in 2007? *Al-Jazeera*, March 3. https://www.aljazeera.com/features/2013/3/3/kenya-what-went-wrong-in-2007 Accessed February 28, 2022.
Casal Bértoa, F. (2021). Database on WHO GOVERNS in Europe and beyond, PSGo. School of Politics & International Relations, Nottingham University. https://whogoverns.eu Accessed December 6, 2021.
Chadwick, A. (2013). *The Hybrid Media System*. New York, NY: Oxford University Press.
Cheng, Y. S. (2007). Introduction: Hong Kong since its return to China: a lost decade? In: *The Hong Kong special administrative region in its first decade*. Hong Kong: City University of Hong Kong Press. 1–48.
Chomsky, N. (1987). *Generative Grammar: Its Basis, Development and Prospects*. Kyoto: Kyoto University of Foreign Studies.
Clark, M. (2021). Research Cannot Be the Justification for Compromising People's Privacy. Facebook/Meta corporation press release. August 3. https://about.fb.com/news/2021/08/research-cannot-be-the-justification-for-compromising-peoples-privacy/ Accessed December 1, 2021.
Cocco, F. (2018). New York Times profits jump 66% amid 'Trump bump'. *Financial Times*, May 3. https://www.ft.com/content/62b5a766-4ece-11e8-a7a9-37318e776bab Accessed February 28, 2022.

CommonCrawl (2021). Distribution of Languages. Statistics of Common Crawl Monthly Archives. Updated monthly. Latest crawl: CC-MAIN-2021-43. https://commoncrawl.github.io/cc-crawl-statistics/plots/languages Accessed December 1, 2021.

Constine, J. (2015). Facebook Finally Lets Its Firehose Be Tapped For Marketing Insights Thanks To DataSift. *TechCrunch*, March 10. https://techcrunch.com/2015/03/10/facebook-topic-data Accessed December 1, 2021.

Curran, J., & J. Seaton (2018). *Power Without Responsibility: Press, Broadcasting and the Internet in Britain*. Eighth ed. Milton Park & New York, NY: Routledge.

Dahlberg, S.; Axelsson, S.; Holmberg, S. (2020). Democracy in context: using a distributional semantic model to study differences in the usage of democracy across languages and countries. *Zeitschrift für Vergleichende Politikwissenschaft*, 14: 425–459.

Dahlberg, S.; H. Kjölstad; A. Ryan (2021a). *GDPR – implementering och konsekvenser för samhällsvetenskaplig forskning*. Executive report, Mittuniversitetet (Mid Sweden University). https://cors.se/wp-content/uploads/2021/05/GDPR-implementering-och-konsekvenser-for-samhallsvetenskaplig-forskning.pdf Accessed December 1, 2021.

Dahlberg, S. et al. (2021b). A Distributional Semantic Online Lexicon for Linguistic Explorations of Societies. *Social Science Computer Review*. In review, manuscript ID SSCR-21-0069.

Daniller, A.; D. Allen; A. Tallevi; D.C. Mutz (2017). Measuring trust in the press in a changing media environment. *Communication Methods and Measures*, 11(1): 76–85.

Davies, M. (n.d.). Representativity. Personal webpage. http://davies-linguistics.byu.edu/ling485/assignments/representativity.asp Accessed June 5, 2020.

De Bolla, P. et al. (2019). Distributional Concept Analysis. A Computational Model for History of Concepts. *Contributions to the History of Concepts*, 14(1): 66–92.

Deutsche Welle (2019). Egyptian independent media outlet Mada Masr says police raided office. *DW online*, November 24. https://www.dw.com/en/egyptian-independent-media-outlet-mada-masr-says-police-raided-office/a-51391162 Accessed February 28, 2022.

Economist (2016). Criminal justice in Mexico: Trials and errors. June 18. https://www.economist.com/the-americas/2016/06/18/trials-and-errors Accessed May 4, 2021.

Elkus, A. (2015). You Can't Handle the (Algorithmic) Truth. *Slate*, May 20. https://slate.com/technology/2015/05/algorithms-arent-responsible-for-the-cruelties-of-bureaucracy.html Accessed December 1, 2021.

European Commission (n.d.). What is a data controller or a data processor? https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/obligations/controller-processor/what-data-controller-or-data-processor_en Accessed December 1, 2021.
Farris, P.W.; N.T. Bendle; P.E. Pfeifer; D.J. Reibstein (2010). *Marketing Metrics: The Definitive Guide to Measuring Marketing Performance*. Upper Saddle River, NJ: Pearson Education.
Fernandez, J. (2010). The East-West divide of Malaysian media. *Malaysian Mirror*, 9 September 2010. http://www.malaysianmirror.com/featuredetail/140-sabah/49237-the-east-west-divide-of-malaysian-media Accessed February 15, 2021.
Fettling, D. (2018). Why no-one speaks Indonesia's language. *BBC Travel*, July 4th. https://www.bbc.com/travel/article/20180703-why-no-one-speaks-indonesias-language Accessed February 28, 2022.
Freedom House (2017). Freedom in the World 2017 – Mexico. https://freedomhouse.org/country/mexico/freedom-world/2017 Accessed May 3, 2021.
Freedom House (2018) Freedom on the Net 2018 – Hungary. https://freedomhouse.org/country/hungary/freedom-net/2018 Accessed May 26, 2020.
Freedom House (2020). Freedom in the World 2020. Accessed May 3, 2021.
Freedom House (2021). Freedom in the World 2021. Accessed May 3, 2021.
Frost, A.A. (2019). Why the Left Can't Stand The New York Times. *Columbia Journalism Review*, Winter 2019. https://www.cjr.org/special_report/why-the-left-cant-stand-the-new-york-times.php Accessed February 28, 2022.
Gold, N. (2020). *Using Twitter Data in Research: Guidance for Researchers and Ethics Reviewers*. Version 1.0, July 16. https://www.ucl.ac.uk/data-protection/sites/data-protection/files/using-twitter-research-v1.0.pdf Accessed December 1, 2021.
Gray, B.; J. Egbert; D. Biber (2017). Exploring methods for evaluating corpus representativeness. Paper presented at the Corpus Linguistics International Conference 2017. Birmingham, UK.
Grzesiek, H. (2021). TikTok Analytics is now available at Quintly. January 14. https://www.quintly.com/product-news/tiktok-data-is-now-available-at-quintly Accessed December 1, 2021.
The Guardian (2020). The Guardian at a glance. Factsheet https://s3.eu-west-2.amazonaws.com/s3-guardian-igad/files/A4_Guardian-at-a-Glance_2020-1.pdf Accessed February 28, 2022.
Halliday, M.A.K. (2005). Computational and quantitative studies. In: J.J. Webster (Ed.) *The collected works of M.A.K. Halliday*, Volume 6. Hong Kong: Continuum.

Helmond, A.; D.B. Nieborg & F.N. van der Vlist (2019). Facebook's evolution: development of a platform-as-infrastructure. *Internet Histories*, 3(2): 123–146. DOI: 10.1080/24701475.2019.1593667

Hindman, M. (2018). *The Internet Trap: How the Digital Economy Builds Monopolies and Undermines Democracy*. Princeton, NJ: Princeton University Press.

High Scalability (2011). DataSift Architecture: Realtime Datamining At 120,000 Tweets Per Second. High Scalability blog. November 29. http://highscalability.com/blog/2011/11/29/datasift-architecture-realtime-datamining-at-120000-tweets-p.html Accessed December 1, 2021.

Hopkins, V. (2021). Hungary media freedom fears mount as broadcaster goes off air. *Financial Times*, February 13. https://www.ft.com/content/3f01c295-ae4e-42f4-8b81-8279ba90cd82 Accessed February 28, 2022.

Human Rights Watch (2017). *Online and On All Fronts: Russia's Assault on Freedom of Expression*. Special report, July 18. https://www.hrw.org/report/2017/07/18/online-and-all-fronts/russias-assault-freedom-expression Accessed November 26, 2021.

Huerta, W.E., & R. Gómez (2013). Concentración y diversidad de los medios de comunicacion y las telecomunicaciones en México. *Comunicación y Sociedad*, 19: 113–152.

Humán Platform (2020). *Hungary turns its back on Europe: Dismantling culture, education, science, and the media in Hungary, 2010–2019*. Oktatói Hálózat – Hungarian network of academics. ISBN 786150 073736.

Industry Arabic (2020). Ranked: The Most Influential Arabic Newspapers (2020 Edition). Industry Arabic. https://industryarabic.com/arabic-newspapers/ Accessed November 25, 2021.

Intellectual Property Office (2019). University and business collaboration agreements: Lambert Toolkit. Published October 6, 2016, updated April 3, 2019. https://www.gov.uk/guidance/university-and-business-collaboration-agreements-lambert-toolkit Accessed December 6, 2021.

Jamnik, M. R., & D.J. Lane (2017). The use of Reddit as an inexpensive source for high-quality data. *Practical Assessment, Research & Evaluation*, 22(1): 1–10.

Jenkins, H.; S. Ford; J. Green (2013). *Spreadable Media: Creating value and meaning in a networked culture*. New York, NY: New York University Press.

Jin, D.Y. (2015). *Digital Platforms, Imperialism and Political Culture*. London & New York, NY: Routledge.

Karidi, M. (2018). News Media Logic on the Move? *Journalism Studies*, 19(9): 1237–1256. DOI: 10.1080/1461670X.2016.1266281

Karombo, T. (2021). South Africa goes after social media as it cracks down on looting and protests. *Quartz Africa*, July 14. https://qz.com/africa/2033328/south-africa-to-monitor-social-media-as-protests-rock-the-country/ Accessed January 22, 2022.

Kayser-Bril, N. (2021). AlgorithmWatch forced to shut down Instagram monitoring project after threats from Facebook. AlgorithmWatch press release, n.d. https://algorithmwatch.org/en//instagram-research-shut-down-by-facebook Accessed December 1, 2021.

King, G., & N. Persily (2020). A new model for industry–academic partnerships. *PS: Political Science & Politics*, 53(4): 703–709. DOI: 10.1017/S1049096519001021

Kitchin, R., & G. McArdle (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*. DOI: 10.1177/2053951716631130

Krippendorff, K. (2013). *Content Analysis: An Introduction to Its Methodology*. Third ed. Thousand Oaks, CA, London, New Delhi, Singapore: SAGE.

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In: M. Hundt; N. Nesselhauf; C. Biewer (Eds.) *Corpus Linguistics and the Web*. Amsterdam: Rodopi. 133–149.

Levine, S. (2021). Letter from Acting Director of the Bureau of Consumer Protection Samuel Levine to Facebook. Federal Trade Commission. August 5. https://www.ftc.gov/news-events/blogs/consumer-blog/2021/08/letter-acting-director-bureau-consumer-protection-samuel Accessed December 1, 2021.

Lewis, S. C.; R. Zamith; A. Hermida (2013). Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods. *Journal of Broadcasting & Electronic Media*, 57(1): 34–52. DOI:10.1080/08838151.2012.761702

Mazzoleni, G. (2008). Media Logic. In: W. Donsbach (Ed.) *The International Encyclopedia of Communication*. Malden: Blackwell Publishing. 2930–2932.

McEnery, T.; R. Xiao; Y. Tono (2006). *Corpus-based language studies: An advanced resource book*. London & New York, NY: Routledge.

Media Ownership Monitor (2020). Intervozes (Coletivo Brasil de Comunicação Social) / Reporters Without Borders. https://brazil.mom-rsf.org/en/ Accessed December 1, 2021.

Mellon, J., & C. Prosser (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, 4(3). DOI: 10.1177/2053168017720008

Moe, H., & A.O. Larsson (2012). Methodological and ethical challenges associated with large-scale analyses of online political communication. *Nordicom Review*, 33(1): 117–124.

Mwita, C. (2021). *The Kenya Media Assessment 2021*. Internews Research Report. March. https://internews.org/wp-content/uploads/2021/04/KMAReport_Final_20210325.pdf Accessed February 28, 2022.

Ndavula, J. (2020). How social media are levelling Kenya's political field – and lessons learnt. *The Conversation*, September 1. https://theconversation.com/how-social-media-are-levelling-kenyas-political-field-and-lessons-learnt-144697 Accessed January 22, 2022.

Nelson, A. (2019). Statement from Social Science Research Council President Alondra Nelson on the Social Media and Democracy Research Grants Program. Official statement, Social Science Research Council. August 27. https://www.ssrc.org/programs/social-data-initiative/social-media-and-democracy-research-grants/statement-from-social-science-research-council-president-alondra-nelson-on-the-social-media-and-democracy-research-grants-program/ Accessed December 1, 2021.

Newman, M.E.J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5): 323–351. DOI 10.1080/00107510500052444

Newman, N.; R. Fletcher; A. Schulz; S. Andı; C.T. Robertson; R. Kleis Nielsen (2021). *Reuters Institute 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 Accessed February 28, 2022.

Nicholson, B. (2022). These were the top publishers on Facebook in January 2022. Newswhip, February 8. https://www.newswhip.com/2022/02/top-publishers-facebook-january-2022/ Accessed February 28, 2022.

Nyanjom, O. (2012). Factually True, Legally Untrue: Political Media Ownership in Kenya. Internews Research Paper. January. DOI: 10.13140/RG.2.1.1933.4485

OECD (2013). Data Provider. Glossary of Statistical Terms. Document created January 29, 2004; updated June 14, 2013. https://stats.oecd.org/glossary/detail.asp?ID=6112 Accessed December 1, 2021.

Palmer, R.; B. Toff; R. Kleis Nielsen (2020). 'The Media Covers Up a Lot of Things': Watchdog Ideals Meet Folk Theories of Journalism. *Journalism Studies*, 21(14): 1973–1989. DOI: 10.1080/1461670X.2020.1808516

PAMCo (2018). PAMCo FAQs. PAMCo – Audience Measurement for Publishers. April 19. https://magnetic.media/news-views/news/pamco-faqs Accessed February 28, 2022.

PAMCo (n.d.) Total Brand Reach (TBR). https://pamco.co.uk/access-to-pamco/total-brand-reach-rules/ Accessed February 28, 2022.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA, & London: Harvard University Press.
Pew Research Center (2021). Newspapers Fact Sheet. June 29. https://www.pewresearch.org/journalism/fact-sheet/newspapers/ Accessed February 28, 2022.
Pratiwi, I.Y.R.; R.A. Asmara; F. Rahutomo (2017). Study of hoax news detection using naïve bayes classifier in Indonesian language. 11th International Conference on Information & Communication Technology and System (ICTS), Surabaya, 2017, pp. 73–78. DOI: 10.1109/ICTS.2017.8265649
Rahutomo, F.; I. Yanuar; R. Andrie Asmara (2018). Indonesian Hoax News Detection Dataset. *Mendeley Data*, V1. DOI: 10.17632/p3hfgr5j3m.1
Raineri, S., & C. Debras (2019). Corpora and Representativeness: Where to go from now? *CogniTextes*, 19. DOI: 10.4000/cognitextes.1311
Reelforge & TIFA (2019). *Kenya Media Landscape Report*. http://www.tifaresearch.com/wp-content/uploads/2019/07/Media-Landscape-in-Kenya-2019-Report-Reelforge-and-TIFA-10.07.2019.pdf Accessed February 15, 2021.
Reporters Without Borders (2020). 2020 World Press Freedom Index. https://rsf.org/en/2020-world-press-freedom-index-entering-decisive-decade-journalism-exacerbated-coronavirus Accessed February 28, 2022.
Reporters Without Borders (2021). Brazil: Climate of hate and suspicion fed by Bolsonaro. https://rsf.org/en/brazil Accessed September 20, 2021.
Riffe, D.; S.R. Lacy; F. Fico (2014). *Analyzing media messages: Using quantitative content analysis in research*. New York, NY: Routledge.
Rivera, I. (2019). RedditExctractoR: Reddit data extraction toolkit. https://cran.r-project.org/web/packages/RedditExtractoR/index.html Accessed December 1, 2021.
Roberson, C. (2020). The Confusing Truth About Corporate Social Media Monitoring. *Business 2 Community*. August 10. https://www.business2community.com/social-media/the-confusing-truth-about-corporate-social-media-monitoring-02334814 Accessed December 1, 2021.
Roberts, D. (2018). The real problem with the New York Times op-ed page: it's not honest about US conservatism. *Vox*, March 15. https://www.vox.com/policy-and-politics/2018/3/15/17113176/new-york-times-opinion-page-conservatism Accessed February 28, 2022.
Rockcontent (n.d.). Pageviews vs Unique Pageviews. https://help.rockcontent.com/en/pageviews-vs-unique-pageviews (Previously: https://help.scribblelive.com/hc/en-us/articles/213583643-Pageviews-vs-Unique-Pageviews) Accessed December 1, 2021.
Rogers, R. (2019). *Digital Methods*. Boston, MA: MIT Press.

Romero, S. (2016). Dilma Rousseff Is Ousted as Brazil's President in Impeachment Vote. *New York Times*, 31 August. https://www.nytimes.com/2016/09/01/world/americas/brazil-dilma-rousseff-impeached-removed-president.html Accessed September 20, 2021.
Roose, K. (2021). Inside Facebook's Data Wars. *New York Times*, 14 July. https://www.nytimes.com/2021/07/14/technology/facebook-data.html Accessed December 1, 2021
Roth, A. (2021). Russian court orders closure of country's oldest human rights group. *The Guardian*, December 28. https://www.theguardian.com/world/2021/dec/28/russian-court-memorial-human-rights-group-closure Accessed February 28, 2022.
Ruths, D., & J. Pfeffer (2014). Social Media for Large Studies of Behavior. *Science*, 346(6213): 1063–1064.
Salaverría, R. & B. Gómez Baceiredo (n.d.). Media Landscapes: Spain. European Journalism Centre. https://medialandscapes.org/country/spain Accessed November 26, 2021.
Shotter, J. (2021). European values: Poland's media fears a crackdown. *Financial Times*, February 22. https://www.ft.com/content/7cc4647c-6781-4566-b0f2-953f17963785 Accessed February 28, 2022.
Silver, L.; A. Smith; C. Johnson; J. Jiang; M. Anderson; L. Rainie (2019). Mobile Connectivity in Emerging Economies. Pew Research Center. March 7. https://www.pewresearch.org/internet/2019/03/07/mobile-connectivity-in-emerging-economies/ Accessed December 6, 2021.
SimilarWeb (n.d. a). Similarweb Data Methodology. https://support.similarweb.com/hc/en-us/articles/360001631538-Similarweb-Data-Methodology Accessed December 1, 2021.
SimilarWeb (n.d. b). Similarweb Vs. Direct Measurement. https://support.similarweb.com/hc/en-us/articles/360002329778-Similarweb-Vs-Direct-Measurement Accessed December 1, 2021.
SimilarWeb (n.d. c). Daily Unique Visitors. https://support.similarweb.com/hc/en-us/articles/360006849237 Accessed December 1, 2021.
Social Science One (2019). Public statement from the Co-Chairs and European Advisory Committee of Social Science One. Statement from the European Advisory Committee for Social Science One. December 11. https://socialscience.one/blog/public-statement-european-advisory-committee-social-science-one Accessed December 1, 2021.
Strömbäck, J. (2021). Media Trust in Europe: Breaking News and Polarized Views. European Liberal Forum Policy Paper No 4. October.
Statistics Estonia (2021). Information and communication technologies. https://www.stat.ee/en/find-statistics/statistics-theme/technology-innovation-and-rd/information-and-communication Accessed November 26, 2021.

Stolyarov, G. (2021). Russia says Twitter mobile slowdown to remain until all banned content is removed. *Reuters*, November 29. https://www.reuters.com/world/europe/russia-says-twitter-mobile-slowdown-remain-until-all-banned-content-is-removed-2021-11-29/ Accessed February 28, 2022.
Sugiura, L.; R. Wiles; C. Pope (2017). Ethical challenges in online research: Public/private perceptions. *Research Ethics*, 13(3–4), 184–199.
Tan, E. (2018). Pamco resets audience measurement with 'total brand reach' for publishers. *Campaign*, April 19. https://www.campaignlive.co.uk/article/pamco-resets-audience-measurement-total-brand-reach-publishers/1462453 Accessed February 28, 2022.
Thomas, R., & S. Cushion (2019). Towards an Institutional News Logic of Digital Native News Media? A Case Study of BuzzFeed's Reporting During the 2015 and 2017 UK General Election Campaigns. *Digital Journalism*, 7(10): 1328–1345. DOI: 10.1080/21670811.2019.1661262
Tiwana, A. (2014). *Platform ecosystems: Aligning architecture, governance, and strategy*. Waltham, MA: Kaufmann.
Treré, E. (2018). From digital activism to algorithmic resistance. In: G. Meikle (Ed.) *The Routledge companion to media and activism*. London & New York, NY: Routledge. 367–375.
Treisman, D. (2011). *The Return: Russia's Journey from Gorbachev to Medvedev*. New York, NY: Free Press.
Tornes, A., & L. Trujillo (2021). Enabling the future of academic research with the Twitter API. Twitter Developer Platform Blog. January 26. https://blog.twitter.com/developer/en_us/topics/tools/2021/enabling-the-future-of-academic-research-with-the-twitter-api Accessed December 1, 2021.
Tuckman, J. (2012). Mexican Media Scandal: Secretive Televisa Unit Promoted PRI Candidate. *The Guardian*, June 26. https://www.theguardian.com/world/2012/jun/26/mexican-media-scandal-televisa-pri-nieto Accessed February 28, 2022.
Twitter Developer Platform (2021). Developer Policy. https://developer.twitter.com/en/developer-terms/policy Accessed December 1, 2021.
Valente, J., & M. Pita (2018). *Digital Monopolies: Diversity and Concentration on Internet in Brazil*. May. São Paulo: Intervozes (Coletivo Brasil de Comunicação Social) / Ford Foundation.
van der Vlist, F.N., & A. Helmond (2021). How partners mediate platform power: Mapping business and data partnerships in the social media ecosystem. *Big Data & Society*. January 2021. DOI: 10.1177/20539517211025061
van Dijck, J., & T. Poell (2013). Understanding Social Media Logic. *Media and Communication*, 1(1): 2–14.
Webhose (n.d.). About Webhose.io. https://webhose.io/about-us/ Accessed September 6, 2021.

Weiss, G., & R. Wodak (2002). *Critical Discourse Analysis: Theory and Interdisciplinarity*. London: Palgrave Macmillan.

Welsh, B. (2020). Malaysia's Political Polarization: Race, Religion, and Reform. In: T. Carothers & A. O'Donohue (Eds.) *Political Polarization in South and Southeast Asia: Old Divisions, New Dangers*. Carnegie Endowment for International Peace. 41–52.

Wiśniewska, K. (2019). Prawicowe tytuły z największymi przychodami z reklam od państwowych spółek. *Press*, March 13. www.press.pl/tresc/56634,prawicowe-tytuly-z-najwiekszymi-przychodami-z-reklam-od-panstwowych-spolek Accessed February 28, 2022.

Wodak, R. (2018). Vom Rand in die Mitte – 'Schamlose Normalisierung'. *Politische Vierteljahresschrift*, 59(2): 323–335.

Zamith, R., & S.C. Lewis (2015). Content Analysis and the Algorithmic Coder: What Computational Social Science Means for Traditional Modes of Media Analysis. *Annals of the American Academy of Political and Social Science*, 659: May. 307–318. DOI: 10.1177/0002716215570576