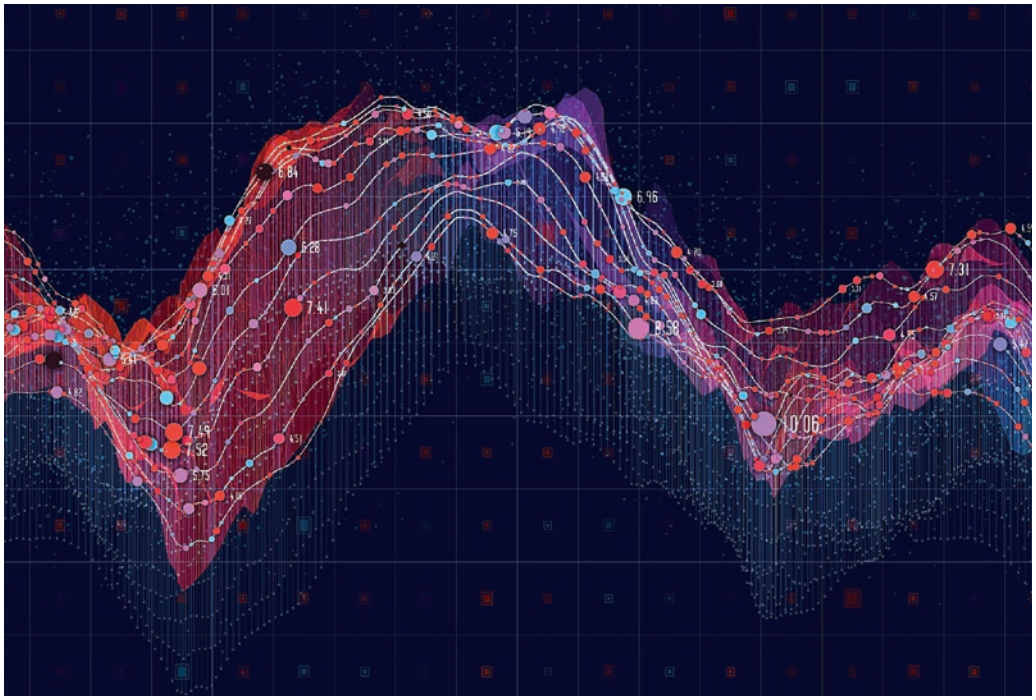
Big Data, Not Big Brother:

New Data Protection Laws and the Implications for Independent Media Around the World

AYDEN FÉRDELINE

June 2019



CIMA
CENTER FOR INTERNATIONAL MEDIA ASSISTANCE

NED
National Endowment for Democracy
Supporting freedom around the world

CIMA
CENTER FOR INTERNATIONAL MEDIA ASSISTANCE

Big Data, Not Big Brother:

New Data Protection Laws and the Implications for Independent Media Around the World

JUNE 2019

Contents

ABOUT CIMA

The **Center for International Media Assistance (CIMA),** at the National Endowment for Democracy, works to strengthen the support, raise the visibility, and improve the effectiveness of independent media development throughout the world. The center provides information, builds networks, conducts research, and highlights the indispensable role independent media play in the creation and development of sustainable democracies. An important aspect of CIMA's work is to research ways to attract additional US private sector interest in and support for international media development.

CIMA convenes working groups, discussions, and panels on a variety of topics in the field of media development and assistance. The center also issues reports and recommendations based on working group discussions and other investigations. These reports aim to provide policymakers, as well as donors and practitioners, with ideas for bolstering the effectiveness of media assistance.

Center for International Media Assistance
National Endowment for Democracy

1025 F STREET, N.W., 8TH FLOOR
WASHINGTON, DC 20004
PHONE: (202) 378-9700
FAX: (202) 378-9407
EMAIL: CIMA@ned.org
URL: https://cima.ned.org

Mark Nelson
SENIOR DIRECTOR

Nicholas Benequista
MANAGING EDITOR

Daniel O'Maley
PUBLICATION EDITOR

National Endowment for Democracy
Supporting freedom around the world

ABOUT THE AUTHOR

Ayden Férdeline is a Technology Policy Fellow with the Mozilla Foundation, where he researches the ongoing development and harmonization of global data protection standards. He previously supported the Internet Society's global public policy team and was a researcher for the data and analytics group YouGov. He is an alumnus of the London School of Economics and is based in Berlin, Germany.

Cover photo: Left side, top; © pixinoo / Shutterstock.com

Introduction

For years, the road to news media financial sustainability was said to be paved with data—digital news outlets were counseled to collect as many details about their readers as possible. Tracking audiences was considered essential for optimizing search engine results, creating content that people want to read, and supporting targeted advertising to fund journalism.

But what started as a way to improve the user experience came with a downside for website viewers: it entailed collecting and processing their personal information, often without their knowledge or consent. Moreover, the drive to collect data has resulted in many independent media outlets in the Global South unknowingly permitting third parties, many of which cannot be identified, to invasively monitor their visitors.

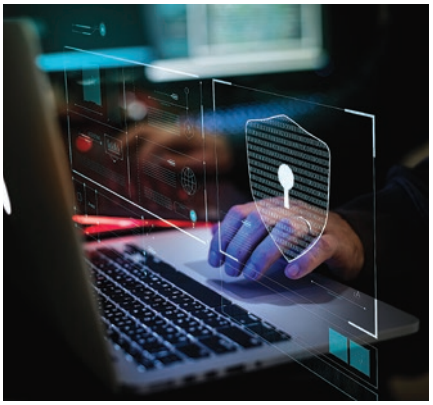
Worldwide, citizens and policymakers are increasingly cognizant of the risks that the burgeoning data economy poses to personal privacy. In recent years, a wave of next-generation data protection laws have emerged that seek to restrict the collection, usage, and sharing of personal information. This is not necessarily a good news story for those news institutions that had successfully harnessed the value of analytics to grow advertising revenue or to better understand their audiences. These data protection regulations have, by design, severely hampered the environment within which many smaller digital media outlets operate. While these laws do not entirely restrict the use of analytics, they do restrict the use of analytic applications that place people at risk of harm. Indeed, there is growing evidence that some of the tracking mechanisms employed by digital news sites are potentially doing just that.

Worldwide, citizens and policymakers are increasingly cognizant of the risks that the burgeoning data economy poses to personal privacy.

To get a better understanding of how new privacy regimes will affect media in the Global South, this paper assembles a new set of findings on the websites of 50 small, independent news publishers from 10 developing countries. It shows that third-party trackers are collecting audience data when people read articles, write comments, send in news tips, and share pieces on social networking platforms. One independent publisher in Nigeria, for instance, had 523 third-party cookies on its homepage collecting audience information. In total, over 150 companies—not all of which could be identified—were found to be invisibly tracking the visitors to these 50 websites. They were collecting IP addresses, which can identify geographic locations, the titles and URLs of news articles read, search queries, and other data. Once collected, this information

could be sold to advertisers or further exchanged with other third parties. It could even reach the hands of governments.

From a media development perspective, the failure of media outlets to protect their visitors against invasive tracking by third parties is troubling for two reasons. First and most importantly, it places the privacy and safety of a publication's readers in jeopardy. Readers need to feel confident that visiting independent news sites, especially those covering sensitive issues, will not put them in danger. Second, from a business perspective, when publishers give away information about their audiences for free, they cede valuable leverage for negotiating with advertisers. In essence, the outsourcing of analytics to third parties potentially puts readers at risk and weakens a site's ability to truly take advantage of the interactions it has with its readers. Taken together, the findings of this study suggest an important new frontier for the media development community and the need to build stronger awareness about and strategies for managing the threats posed by tracking the readers of independent media.



The findings of this study suggest an important new frontier for the media development community and the need to build stronger awareness about and strategies for managing the threats posed by tracking the readers of independent media.

This report also provides an overview of the latest regulatory developments in the data protection field, such as the European Union's General Data Protection Regulation (GDPR). While much of the public debate has been about policy changes in Europe, the impact has been much broader in geographic scope. The changes underway directly impact independent news outlets in many developing countries. This review of new laws is followed by a detailed description of the various trackers currently used on news websites, and the ways that they potentially come into conflict with new data protection laws and regulations. Then, the report analyzes research on web tracking technologies used by news media websites in 10 developing countries, exposing pervasive tracking that ultimately may not benefit either the readers or the news organizations themselves. And finally, it concludes with suggestions about how news organizations and other media development stakeholders might be able to take advantage of the global shift in data protection laws and regulations to strengthen independent media.

DEFINITIONS

What is privacy?

A generally accepted definition of privacy is "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others."[1]

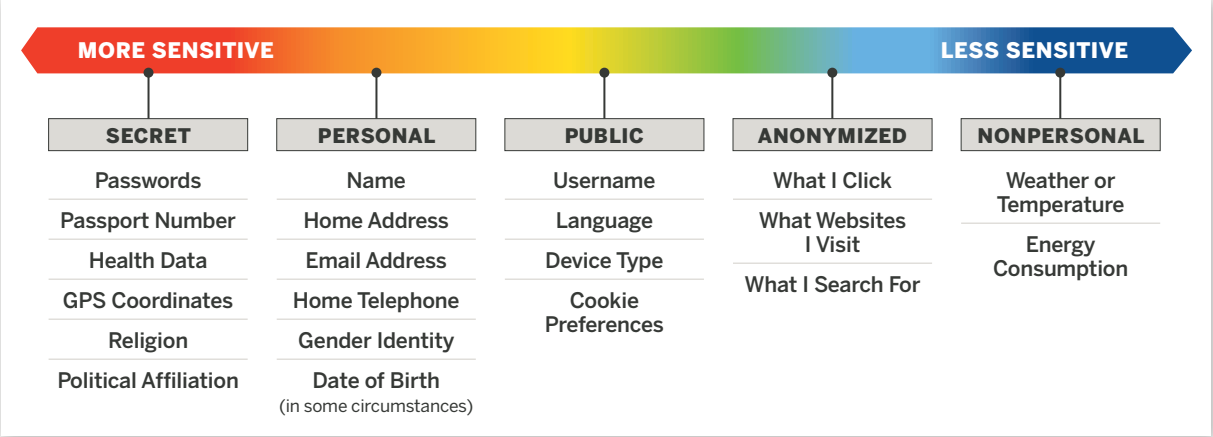
What are data protection laws?

Data protection laws are frameworks that seek to regulate the collection, storage, and processing of information about individuals.

Is all data protected by data protection laws?

No. Data protection laws apply only to personal and sensitive information. Data protection laws do not protect nonpersonal data,[2] anonymized data, or public data.[3] This is an important distinction because many common data analytic practices do not use personal information at all.

FIGURE 1. The Range of Consumer Privacy Levels



Note: While privacy is a disputed concept in law and philosophy, and attitudes toward how personal information is used vary from individual to individual, it is generally accepted that some data elements are more sensitive than others.

What is personal information?

There is no universal definition of what is or is not personal information. However, a common definition found within many national laws and international agreements modelled after the European Union's GDPR is that personal data "means any information relating to an identified or identifiable natural person."[4] Some data elements very clearly count as personal or secret information, such as a name or passport number. But the answer is not so straightforward for other elements. For example, a date of birth in and of itself is not personal information. But if that can be combined with a street address and one's gender, it could be used to identify someone, and in that instance should be treated as personal information.[5]

Cutting Through the Complexity: Privacy, Data Protection, and Personal Information

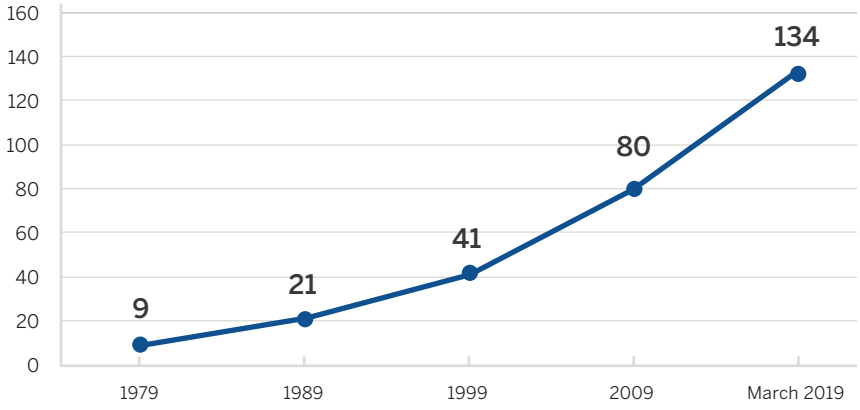
As the internet has grown in social and economic importance, more and more people have begun engaging with technologies that surreptitiously undermine their privacy. Businesses have emerged with business models that are based on gathering, using, and selling personal information without the data subject's knowledge or explicit consent.

Personal information has come to be seen by some companies as an economic asset to be harvested or as a tool to better inform editorial investments.

Media companies have also begun using similar data sets to understand reader preferences, follow up on stories, and create content that responds to audience demand. Personal information has come to be seen by some companies as an economic asset to be harvested or as a tool to better inform editorial decisions. At the same time, individuals have reported feeling powerless to stay in control of how their personal information is being used. Increasingly, however, there are regulatory barriers that restrict these activities.

As of March 2019, 134 countries had enacted data protection laws,[6] while 26 others[7] had drafted legislation with some degree of government support. While there are exemptions within many data protection laws for journalistic activities like newsgathering, there are almost always implications within these laws for the "business side" of media companies

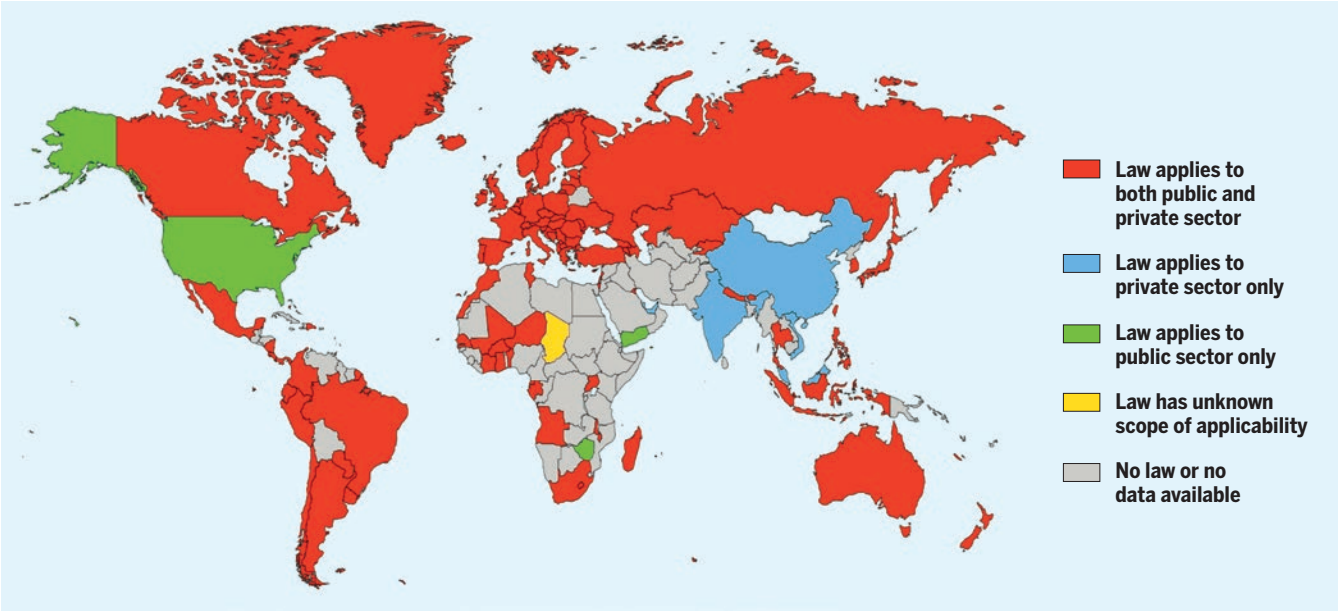
FIGURE 2. Number of Countries with Data Protection Laws



Note: The number of countries with data protection laws has risen to 134.

SOURCE: Graham Greenleaf, "Global Data Privacy Laws 2017: 120 National Data Privacy Laws," Privacy Laws & Business International Report 145 (2017): 10-13, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2993035. Supplemented by further research by the author.

FIGURE 3. Countries with Data Protection Laws



Note: Most data protection laws apply to both the public and private sectors, but there are exceptions. The enforcement of these laws varies from jurisdiction to jurisdiction.

SOURCE: These data are based upon original research and analysis by the author.

and journalistic institutions. These laws are particularly likely to apply if a news organization's website collects data about its audience, loads elements onto a webpage from a third party, or uses cookies. It would not be feasible for most entities to adhere to the unique laws of 134 countries; however, it is also not necessary to do so. Adopting the highest data protection standard available is the most straightforward approach to compliance. At this time, that standard is the European Union's General Data Protection Regulation (GDPR). Achieving compliance with the GDPR and its principles of accountability and data minimization would place most media organizations in good standing, even if they operate in a different regulatory environment.

Although European nations represent a minority of those jurisdictions with data protection laws, Europe continues to have an outsized influence on the development of data protection laws in Africa, Asia, and the Americas. This is unlikely to change in the foreseeable future because the Council of the European Union has advised the European Commission that it cannot negotiate away privacy rights in trade agreements.[8] Countries that wish to trade with the European Union, and, in particular, have data flows with the trading bloc, will thus face pressure to implement data protection laws that are influenced by the European standard.

Although European nations represent a minority of those jurisdictions with data protection laws, Europe continues to have an outsized influence on the development of data protection laws in Africa, Asia, and the Americas.

Key Historical Developments in Privacy Law

Privacy rules and norms that govern action or inaction related to our personal information have been interpreted in a similar way around the world for some time. In 1974 the United States adopted the Privacy Act,[9] a federal law that sought to safeguard information about individuals held by federal agencies. The act codified into law the recommendations developed by an independent advisory committee in 1973 that had analyzed the consequences of using electronic systems to maintain records about people.[10] Their report shaped our contemporary understanding of information privacy and it remains relevant some four decades later. In short, the committee recommended adopting five principles:

Data protection laws are frameworks that seek to regulate the collection, storage, and processing of information about individuals.

1. **There must be no secret record-keeping systems.**
2. **Individuals must be able to find out what information about them is in a record and how it is used.**
3. **Information cannot be obtained for one purpose and then used for another purpose without the consent of the individual concerned.**
4. **There must exist a right to correct inaccurate records.**
5. **Organizations are responsible for ensuring that their record-keeping systems are secure and reliable, and must take precautions to prevent the misuse of data.**

Following the passage and implementation of the act, the United States advocated for these principles internationally. Today, they can be found in every major privacy protection instrument, including the African Union Convention on Cyber Security and Personal Data Protection, the Asia-Pacific Economic Cooperation's Cross-Border Privacy Rules, the Organisation of Eastern Caribbean States' Data Protection Bill, the Organisation for Economic Co-operation and Development's Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, the Council of Europe's Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, the European Union's Data Protection Directive, and the GDPR.

The GDPR came into effect in 2018. It was a major revision to European law that significantly built upon the principles contained within the US Privacy Act of 1974 and the values advanced within the EU Data Protection Directive of 1995, aiming to prohibit the excessive collection, use, and disclosure of personal information without disproportionately impeding

commerce, free expression, or freedom of association. Whether this balance was successfully achieved remains hotly debated, but what is uncontested is that this legislation forced companies around the world to review their data processing activities.

GDPR in a Nutshell

The GDPR codifies into law a risk-based approach to protecting the privacy of natural persons. It requires privacy by design and by default, mandates accountability for data controllers, and grants individuals new rights, including the rights to erasure and to control and transparency over how their personal information will be used. The GDPR states that personal information must be retained for the shortest period of time possible and that there must be limits on who can access it. It also imposes significant restrictions on how and when personal information may be shared with third parties. Further, the GDPR grants new protections to sensitive information like medical data, and Article 7 states that if an individual is asked to consent to a data processing practice, their consent must be a "freely given, specific, and unambiguous" indication of their intent. Most strikingly, the penalties for noncompliance are set at €20 million ($22.4 million) or 4 percent of global revenue, whichever is higher, even where there is no ill intent on the part of the data controller.

The GDPR codifies into law a risk-based approach to protecting the privacy of natural persons. It requires privacy by design and by default, mandates accountability for data controllers, and grants individuals new rights, including the rights to erasure and to control and transparency over how their personal information will be used.

Extraterritoriality

A major difference between the GDPR and other data protection laws is that it has extraterritorial effect, meaning that enforcement is theoretically possible outside of the borders of the European Union. This has made Europe, in the eyes of some, the "world's data police."[11] The consequence of this is that under the GDPR, even organizations outside of the European Union must comply with the GDPR when they process data belonging to individuals in the European Union. Because of the global nature of the internet, it is easy to imagine a European resident who is protected by the GDPR visiting the website of a publisher, say, in Belarus or Mongolia. At least in theory according to European Union regulations, that Belarusian or Mongolian publisher must adhere to the GDPR if collecting analytic data about that European resident. If such extraterritorial enforcement actually happens, there would be profound implications here for the digital media ecosystem worldwide.

At this point in time it is difficult to know what obligations will actually be enforced on entities located outside of the European Union. One of the largest ambiguities that the internet presents, when it comes to the applicability of legislation, is that it is a space where conventional nation-state borders do not exist and where traditional modes of interstate legal cooperation have struggled to keep pace with the realities of a Web 2.0

world. This is why earlier data protection laws have been difficult to enforce and why the European Union, in the GDPR, has sought to make its legislation applicable in all environments. The fear, however, is that this jurisdictional overreach could lead to a legal arms race that could have unpredictable and unintended consequences. So far, the GDPR's enforcement bodies have been reluctant to impose penalties on data controllers outside of the European Union. In one notable example, the United Kingdom's Information Commissioner's Office sent a letter to the *Washington Post* advising that its website did not comply with the GDPR, but it did not take any formal enforcement action.[12] This suggests that, at least for now, the European Union will rely on indirect means of enforcing the GDPR outside of its borders, incentivizing self-compliance through fear of reputational damage.

So far, the GDPR's enforcement bodies have been reluctant to impose penalties on data controllers outside of the European Union.

GDPR Is the New Global Standard

Ten countries outside of Europe have now updated their earlier data protection bills to enact many (or all) of the principles contained within the GDPR. This trend appears to be continuing, with new or updated bills pending in Algeria, Indonesia, Thailand, and Tunisia that appear to have been modelled after the GDPR. Pakistan, which does not have any data protection legislation at present, currently has a bill under consideration that would adopt large chunks of the GDPR.

"I assume that lawmakers just copied and pasted the GDPR and left some things out actually," said Salwa Rana, legal officer at Media Matters for Democracy in Pakistan.[13] "And these things were that you need to inform the data subject of any leak that takes place, that the data subject has the right to be forgotten, and extraterritoriality." Rana said the question of extraterritorial application is one that remains unaddressed. "This was one of the main questions that was raised in one of our consultations: Is enforcement of the law going to be limited to Pakistan? The problem is that we have the federal investigation authority which is going to be responsible for any violations under this law outside of Pakistan, yet under the proposed legislation, they haven't given them any power." While the GDPR's exemptions for journalistic activities remain in the proposed bill, media organizations in Pakistan have not been actively involved in drafting this law. "There hasn't been much response from media companies, but I feel like the way this law is going, they are going to have to begin participating."

Regardless of whether or not more countries adopt the GDPR's provisions, given the global nature of the internet and many businesses' desire to trade with member states of the European Union, a need to comply with the GDPR has incentivized businesses such as Microsoft to voluntarily adopt higher privacy and data protection standards for their entire operations worldwide, even where they are under no legal obligation to do so.[14]

GDPR Implications for the Media Industry

Article 4 (7) of the GDPR defines a data controller as "the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data." By this broad definition, it is difficult to imagine any media organization with either a list of subscribers or a website with analytic functions that *would not* be considered a data controller. As a result, any journalistic institution whose content is accessible to European residents will need to think carefully about how the GDPR may impact their business development activities or editorial functions. Some potential ramifications include the following:

Impacts on Newsgathering

The GDPR *states* that the

> *... processing of personal data solely for journalistic purposes, or for the purposes of academic, artistic or literary expression should be subject to derogations or exemptions from certain provisions of this Regulation if necessary to reconcile the right to the protection of personal data with the right to freedom of expression and information.*[15]

While this language provides the media with significant leeway to be able to publish journalistic work, it does not assist journalists in accessing information for journalistic purposes.

Ioana Avadani, executive director of the Center for Independent Journalism in Bucharest, said the GDPR has been used in Romania to protect those in positions of power.[16] "What we witnessed immediately after the GDPR is that institutions started to invoke the GDPR as a reason not to release information," she said. "They were not keen on releasing information before, so what they got is just another reason, and they are very happy that this is a legal reason to justify their less-than-transparent attitude."

Avadani pointed to an example of a protest that occurred in August 2018 where riot police in Bucharest behaved in a violent manner and physically assaulted demonstrators. After a journalist asked who had authorized this action, the police invoked the GDPR and refused to name the authorizing officer. "It was a clear case of public information, and they still refused to say it because they wanted to protect the government," said Avadani. Their next reaction was to use the GDPR to attempt to force the journalist to reveal their source, claiming the police department had an obligation under the GDPR to investigate a data breach. "It was not a genuine concern for the protection of the police officer's privacy, it was just a way to protect the authorities."

© Urban Fenix / Shutterstock.com

"What we witnessed immediately after the GDPR is that institutions started to invoke the GDPR as a reason not to release information... They were not keen on releasing information before, so what they got is just another reason, and they are very happy that this is a legal reason to justify their less-than-transparent attitude."

— IOANA AVADANI,
Center for Independent Journalism in Bucharest

While it is doubtful such a fine would stand up in the highest courts of the European Union, for smaller media outlets the fear of costly, ongoing litigation could ultimately have a chilling effect on journalism.

This is not the only case of the GDPR being abused in Romania. RISE Project, a non-profit investigative journalism organization, was threatened with a €20 million ($22.4 million) fine from Romania's data protection authority after publishing a post on Facebook that accused a prominent Romanian politician of theft. RISE Project subsequently published a letter it had received from the National Supervisory Authority for Personal Data Processing, which demanded that it disclose within 10 days "how and when RISE Project obtained the information ultimately posted to Facebook, who their source was, how they stored the documents, and what other personal information RISE Project has on [the politician] and their friends,"[17] or face a penalty of €20 million. While it is doubtful such a fine would stand up in the highest courts of the European Union, for smaller media outlets the fear of costly, ongoing litigation could ultimately have a chilling effect on journalism.

"Right to Erasure" Impact

The "right to erasure," also known as the right to be forgotten, has garnered significant attention but is often misunderstood. The right is not absolute, with Article 17 (3) of the GDPR offering a public interest exemption intended to safeguard against predicted abuses. The problem is that the GDPR's Recital 153 *states* that "Member States law should reconcile the rules governing freedom of expression and information, including journalistic, academic, artistic and or literary expression with the right to the protection of personal data pursuant to this Regulation."[18] This means that there could be a patchwork quilt of interpretations for how this article should be implemented. Unfortunately, in Romania, the data protection authority has settled upon a definition that seems to have prioritized the right to privacy over freedom of expression in all circumstances.

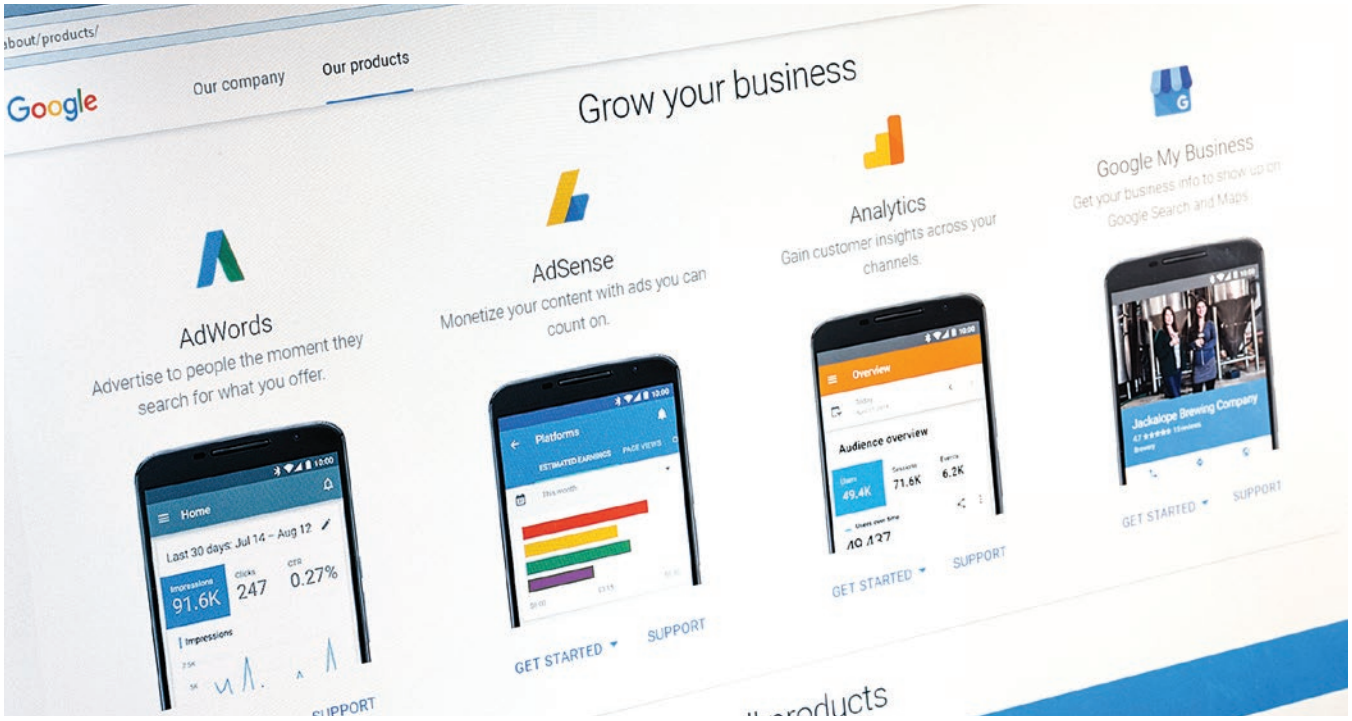
Ziarul de Iasi, a local newspaper in Romania with a circulation of 5,000 copies per week, received a right to erasure request to delete an article from its online archive. After the newspaper refused to remove an article from nine years earlier about a public figure who had engaged in improper behavior, the National Supervisory Authority for Personal Data Processing sent a letter imposing a fine of 3,000 leu ($725) per day until the article was deleted. "For a local newspaper this is huge," said Avadani. "In this particular case, *Ziarul de Iasi* is going to challenge the request in court. However the editor-in-chief told me if he keeps receiving requests like this, he may not be able to afford to keep challenging them."

Impacts on Internal Operations and Website Functionality

Under the GDPR, data controllers are obligated to ensure that both their data processing practices and the data processing practices of third parties comply with the regulation. This necessarily requires that media organizations more closely scrutinize the activities of the third-party vendors they work with.

Ala'a Alzghoul, an information systems specialist with Arab Reporters for Investigative Journalism in Jordan, explained how the GDPR prompted his organization to develop internal procedures for assessing how third parties handle personal information.[19] "For example, we use Google Analytics to collect some data for the user experience. Before we added their plug-in, we first read the privacy policy of Google Analytics and asked for every detail as to what data this tool is collecting, we tracked what they actually do, and we mention those details in our privacy policy." Alzghoul explained that the GDPR also prompted Arab Reporters for Investigative Journalism to develop new procedures for handling personal information. "To prevent any data leakages, we moved from regular databases to encrypted databases," he said. "We have a new policy to protect the personal data that we collect, and to prevent employees [from] just copying the data onto their laptops. But this happened because of the GDPR, not because we were afraid of the laws here in Jordan."

Under the GDPR, data controllers are obligated to ensure that both their data processing practices and the data processing practices of third parties comply with the regulation.



© pixinoo / Shutterstock.com

Understanding Websites and Analytics and Balancing Interests

When someone visits a newsstand and buys a printed newspaper, they receive a complete product. But when a visitor browses a webpage, their web browser does not download one file. Rather, the web browser reads the code, downloads the required content from various sources, and renders the page. This all happens in milliseconds. The output may appear to the reader as one complete package, but more happens behind the scenes than many people realize, with content typically being downloaded from both first-party and third-party sources.

1 First-party content comes from the same location as the webpage itself. For example, if you visit *https://es.MercoPress.com*...

MercoPress.
South Atlantic News Agency

2 When the webpage loads, this image is downloaded from *https://es.mercopress.com/web/img/mp-logo.png*. Because it comes directly from MercoPress's website, it is first-party content.

First-party content does not currently face legal or regulatory challenges.

3 But the ads? They are downloaded from *pagead2.googlesyndication.com/pagead/js/adsbygoogle.js*.

Third-party content comes from a different source. Because this content comes from GoogleSyndication.com and not MercoPress.com, it is third-party content.

Third-party content is impacted by the GDPR.

When content is being downloaded, the browser sends an HTTP request to either retrieve information from a server or send data to a server.[20] As part of this interaction, the server obtains the visitor's IP address to learn who it is interacting with. You could think of an IP address as the return address on a letter you mail; it is a unique number that essentially identifies you by the device you're using to connect to the internet, and can be linked to all the online activity you engage in on that device.

When first-party content is downloaded, a website visitor would reasonably expect that they are sharing their IP address to access that content. The situation becomes murkier with third-party content. Because the website viewer's IP address is being collected by an external source with which they do not have a direct relationship—and since these third-party content elements can either be invisible, blend into the webpage, or just load on the webpage without the individual's explicit consent—their IP address would be collected by a third party without the individual's knowledge or approval.

Moreover, it is rare that only an IP address will be captured. While an IP address does constitute personal information, it is extremely common for third parties to collect information on individuals through cookies, web beacons, and application program interfaces, among other technical measures, as people browse digital properties. These allow for individuals to be targeted in a much more granular manner and to be tracked across the entire internet.

When first-party content is downloaded, a website visitor would reasonably expect that they are sharing their IP address to access that content. The situation becomes murkier with third-party content.

How Website Visitors are Tracked	
Cookie	A cookie is a message that a server sends to a web browser to store on the website visitor's computer. This file is then sent back to the server each time the visitor's web browser requests content from that particular server.
Web beacon	A web beacon is a small image, usually one pixel by one pixel in size, that is discretely placed on a website to monitor visitor behavior. When the image loads, the web beacon passes information along to the server where the image is stored, including the IP address of the computer that retrieved the image, the time the web beacon was viewed, the type of browser that retrieved the image, and cookie values.
Application program interface	An application program interface determines how different software applications and components should interact with others. They are building blocks that website developers can use to pull and share data. For instance, Amazon's Product Advertising application program interface allows another website to search Amazon's product inventory and to then add personalized functions to its website advertising Amazon's products.

There are various fingerprinting algorithms that have enabled data sets to be analyzed in a manner that, for all practical purposes, could uniquely identify an individual with a high degree of accuracy.

Historically it has even been possible for companies to track website viewers without using cookies or web beacons or deploying application program interfaces. There are various fingerprinting algorithms that have enabled data sets to be analyzed in a manner that, for all practical purposes, could uniquely identify an individual with a high degree of accuracy. When a visitor downloads a file from a third party, the third party necessarily obtains a user agent string (the website visitor's operating system, web browser type, and version number) and accepts headers (the type, version, and capabilities of the browser that is making the request so that the server returns compatible data). If JavaScript is enabled, it can communicate the names of browser plug-ins that are installed, and these plug-ins can be called upon to share system-specific attributes. Many of these attributes are, in and of themselves, harmless, but when aggregated, can effectively and easily lead to the identification of a user.[21]

To grasp how the tracking of readers involves a variety of distinct and independent entities, imagine a scenario in which an individual lands on a fictitious new site, NewsWebsite.com, to read an article on nutrition. The reader's presence on the site is collected by Analytics.com, a third-party audience measurement tool firm that NewsWebsite.com has enabled on its website. Analytics.com collects data from millions of websites using cookies, and this data could include the visitor's past shopping behavior, interests, time zone, ethnicity, browser language preferences, and gender, among other information. Based on these data compiled by Analytics.com, another third-party, Shopping.com, knows that the visitor is female, aged 40, and previously spent $60 on a book. Shopping.com could now infer that the visitor is likely to be interested in hardcover recipe books, and so sends a request to Advertising.com to load an advertisement for a hardcover recipe book on the next page that the visitor loads on the NewsWebsites.com site. In this scenario, personal data about the reader would have been circulated with at least four different entities, some of which the reader herself might not even know about.

Over the past three years web browsers and mobile devices have begun masking header information by default. This reduces, but does not entirely eliminate, the potential for reidentification through this manner. These changes have arguably been implemented in reaction to new privacy regulations like the GDPR, which oblige data controllers to provide individuals with an effective means of exercising their data rights. Users with older mobile devices or web browsers that have not been updated may be particularly vulnerable to identification through fingerprinting.

User Tracking by Independent Media Outlets in Developing Countries

This report's analysis of news websites in developing countries sheds light on the pervasiveness of user tracking on these news sites. Of the 50 small, independent publishers studied, 92 percent contained third-party tracking devices such as cookies and web beacons. Most of the tracking devices whose owners we could identify were transmitting data to the United States or the European Union.

However, 15 percent of the tracking devices we found on independent news websites had no easily identifiable ownership. Some of the most pervasive trackers on independent news websites in Kenya, Nigeria, and Ukraine, for example, actively masked their identities. In such cases the average website visitor would not know who is collecting their personal information or for what purpose. At least when the owner of a cookie can be identified, users can make contact with them to exercise their rights.

Overall, the analysis included 100 news websites in Argentina, Brazil, Egypt, India, Indonesia, Kenya, Nigeria, Syria, Ukraine, and Uruguay—five small publishers and five large publishers in each country. The measurements were conducted using the open source OpenWPM platform, which was developed by scientists at Princeton University. This tool has been used in 22 academic studies, and it allows researchers to systematically and reliably quantify, understand, and uncover the ways in which website users are tracked across the measured websites.[22] The OpenWPM tool was deployed using a local virtual private network (VPN) to imitate the website experience of a local website visitor (with the exception of Syria, which does not have a VPN, where sites were visited using a Turkish VPN).

The small, independent publishers chosen for analysis were selected based on recommendations from respected journalists and media policy advisors in the field. The ownership structures and funding sources of these websites were also taken into account to verify their independence. To be included, the sites also had to be posting original content consistently for three or more years. For the large publishers, we included the five most visited news websites in the given country in January 2019, per Alexa Internet's rankings. In Uruguay, the top three most visited news websites were Argentine, so instead the analysis included the five most visited news websites that were published out of Uruguay.

Some of the most pervasive trackers on independent news websites in Kenya, Nigeria, and Ukraine, for example, actively masked their identities. In such cases the average website visitor would not know who is collecting their personal information or for what purpose.

Key Findings

92%
of these independent media websites contained third-party tracking devices

167
unique companies are monitoring website visitors on independent media websites

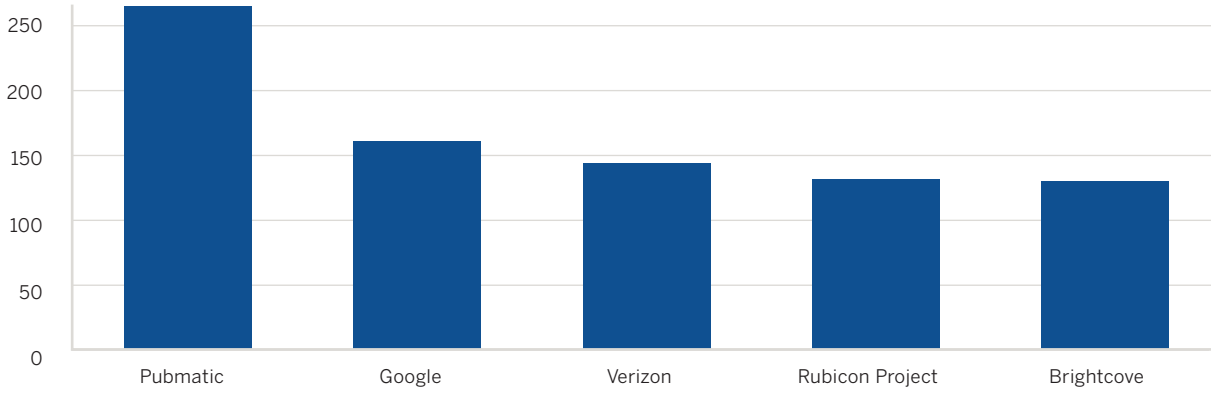
523
cookies were found on one single webpage of an independent news website in Nigeria

1
independent news website in Ukraine had been compromised by a third party and was distributing invasive malware to website visitors

While the study was not exhaustive, it is significant because it found that over 150 companies—not all of whom we could identify—were invisibly tracking the visitors to these 50 independent news websites. They were collecting IP addresses, which can identify one's geographic location, the titles and URLs of news articles read, search queries, and other data. Once collected, this information could be sold to advertisers or further exchanged with other third parties. Any data that are collected are also vulnerable to being stolen in a data breach, or obtained by a government through a court order.

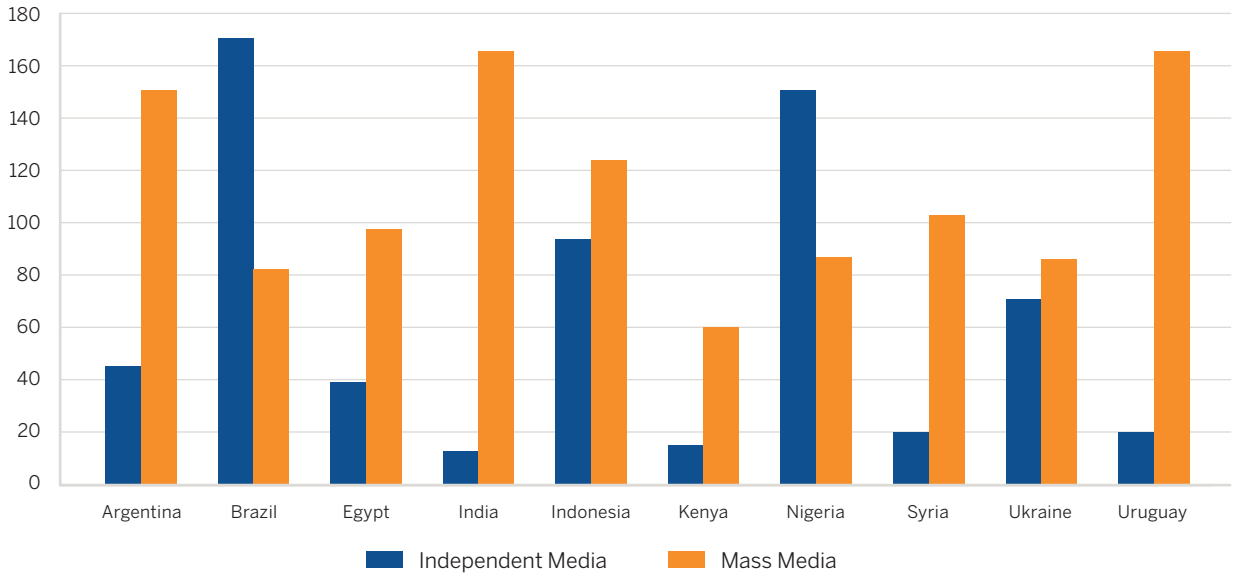
The situation was no better for large, mainstream media websites in the same countries. Overall, 98 percent of the large news websites that we analyzed contained third-party cookies. With the exception of websites reviewed from Brazil and Nigeria, the large news websites contained more third-party cookies than their independent counterparts did.

FIGURE 4. Top Five Identifiable Tracking Companies on Independent Media Websites



Note: The top five tracking companies across the 50 independent media websites whose ownership could be identified, as tested on April 11, 2019. This excludes trackers—some of which would have otherwise made the top five—that operate in an opaque manner and do not disclose for whom they are collecting and sending data.

FIGURE 5. Number of Third-Party Trackers (Average on Website Homepage)



Note: With the exception of Brazil and Nigeria, the homepages of large publishers tended to contain more third-party trackers than the homepages of small publishers. Test conducted April 11, 2019.

Uses and Purposes of Tracking Devices

In the analysis of 50 independent news websites, seven common uses of tracking devices were identified—along with one uncommon, but problematic, use.

Commons Uses of Tracking Devices (in alphabetical order)	
Advertising	Displaying online advertisements creates a very significant stream of income for many news websites. Tracking devices are commonly used to embed third-party advertisements and to exchange reader data to display targeted, behavioral advertisements. The most common advertising networks we found were PubMatic, Google AdSense and Google DoubleClick, and Rubicon Project. **Potential for Privacy Violation: HIGH**
Audience Measurement	News websites deploy technical measures to determine the number of unique website visitors, the number of pages visited, and the average time spent on the webpage. Cookies are used to determine repeat visitors and data can be exchanged with third parties to build demographic profiles of visitors. The most common audience measurement instruments we found were Google Analytics and Adobe Experience Cloud. **Potential for Privacy Violation: MEDIUM**

continues next page

Commons Uses of Tracking Devices (in alphabetical order)

Content Hosting	Most news websites use third-party hosting providers, either to host their website or to cache content so that it loads more quickly. Common providers include Amazon Web Services and Cloudflare. Some publishers use third-party content libraries, such as Adobe Fonts, to improve the appearance of their websites. Others use tools like YouTube to host videos because it is either cheaper than self-hosting that content, or easier to extend the functionality of their websites by turning to a third-party application. In all these instances, when content is being loaded through an intermediary, it exposes a visitor's IP address (or more) to the third party. **Potential for Privacy Violation:** MEDIUM
Design Optimization	News websites sometimes conduct design experiments using real audiences to understand what design changes could keep visitors on their website longer or improve website usability. These tools, like Apptimize, Optimizely, and Splitforce, do not typically collect personal information and data are usually accessible only by the publisher. **Potential for Privacy Violation:** LOW
Recommendation Systems	Some news websites use content recommendation engines to encourage website viewers to remain on the same website, or to visit a partner's website to read an article that the publisher believes the reader will find interesting. These tools are thought to be behind many of the opaque tracking devices that we found, as their content changes dynamically and is updated by a code that the website owner is unable to closely scrutinize. **Potential for Privacy Violation:** HIGH
Social Media Share Buttons	Many news websites have embedded social media share buttons into their websites. These are intended to facilitate sharing news articles via the most common social media platforms. The most common social media share buttons we saw were for Facebook, which, if its button is embedded into a webpage, results in audience information being shared with Facebook. Other buttons included the Twitter icon and the ShareThis widget. **Potential for Privacy Violation:** HIGH
Visualizations	Some news websites use third-party tools to attractively display their stories, for instance, by embedding interactive maps, using Scribd to embed annotated documents into a page, or using Tableau to create column graphs or pie charts. **Potential for Privacy Violation:** MEDIUM

Uncommon Uses of Tracking Devices

Malware Distribution	One independent Ukrainian news website was identified as a vector for disseminating malware. Malware is software that is designed to allow an unauthorized actor to gain access to a computer. It was difficult to decisively attribute the source and creator of the malware being distributed, as it actively masked its identity, but this particular content was embedded into the Ukrainian website through a third-party advertisement and appeared to be transmitting data to servers in Russia. **Potential for Privacy Violation:** HIGH

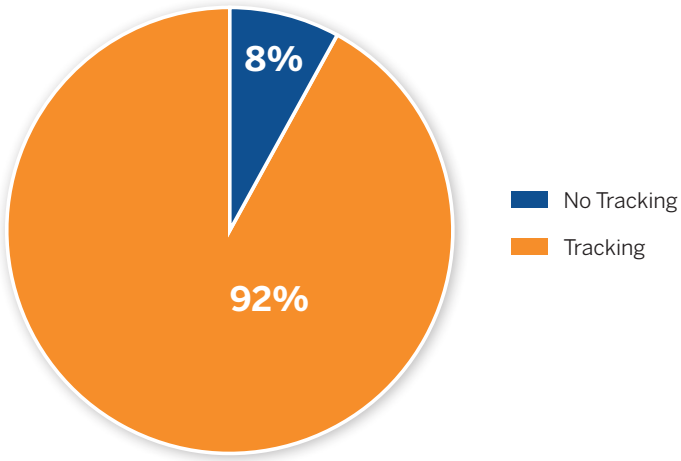
Ninety-two percent of the independent media websites and 98 percent of the mass media websites reviewed contained third-party trackers of some kind. In the context of independent media websites, these trackers were most commonly deployed for audience measurement purposes, followed by advertising and marketing purposes, and then by the inclusion of social media share buttons.

According to Valentina Pavel, who researches data ownership at Privacy International, "user tracking and exploitation of data is still the default for news websites, but this is changing."[23] She said readers are looking for something else and that publishers can turn data protection principles to their competitive advantage. "Be fair and clear to your readers, show them you have been thoughtful about the way you are handling their data, and collect only the type of data that is necessary for the smooth running of the website and explain in plain language why you made those choices." She noted that large publishers like the *New York Times* have dropped behavioral advertisements from their websites altogether without suffering any revenue impact and believes this has paved the way for smaller publications to do the same. "People are looking for real guarantees that their data is not going to be exploited, so by all means, don't sell or share user data, and stop or limit using first- or third-party cookies," said Pavel. "If others do it, why can't you?"

"Be fair and clear to your readers, show them you have been thoughtful about the way you are handling their data, and collect only the type of data that is necessary for the smooth running of the website and explain in plain language why you made those choices."

— VALENTINA PAVEL,
Privacy International

FIGURE 6. Third-Party Cookies on Homepages of Independent Media Websites



Note: It is important to note that during our test, we did not press the "consent" button on websites to permit the collection and use of cookies. Accordingly, these results should be read to show that 92 percent of websites tested had third-party tracking devices that activated even in the absence of visitor consent.

Immediate Privacy Gains Are Possible

Around the world, data protection laws and regulations are changing the digital ecosystem; by extension, they are changing the online publishing world too. Coming into compliance with new data protection laws like the GDPR and other applicable regimes can be difficult, and that appears to be a leading reason that some website owners have not taken action.

One simpler mechanism of coming into compliance with the GDPR is to migrate away from using third-party tools, services, and applications and to instead use self-hosted tools. This immediately limits exposure to the data collection and processing practices of third parties, and grants readers a greater degree of privacy protection, as information that could lead to their identification will no longer be circulating outside of the websites they're visiting.

Moving from Third Party to First Party (in alphabetical order)	
Advertising	Using a third-party ad exchange, like Google AdSense, to fund journalism sites makes it impossible to eliminate the privacy risks that these tools present. However, a website owner may decide not to use an ad exchange, opting instead to handle advertising sales internally. This has been an approach that larger publishers have been taking post-GDPR, as it can also result in higher revenue from direct sales negotiations. If advertisements are not targeted to the individual website viewer and are instead of a general nature, this is fairly simple to resolve. Provided no personal information about the website viewers is exchanged with the advertiser, no privacy violations can occur. Much, if not all, of the information typically found in a media kit about audience demographics does not constitute personal information. If personal information is to be exchanged with third parties, a careful assessment will need to be conducted before this happens. This assessment must consider both the fundamental rights and freedoms of the individuals concerned, their reasonable expectations for how and why their personal information would be used, and the grounds for why and how the third party would reasonably use those data. **Ease of Migration:** **DIFFICULT**
Audience Measurement	There are audience measurement tools that can be self-hosted, such as Cryptolog, which allows both internal analytical logging and total control over how long data are retained. However, these tools are not as easy to use as third-party ones and are not as rich in features. Significant training may be required to learn how to use them effectively. **Ease of Migration:** **DIFFICULT**

continues next page

Moving from Third Party to First Party (in alphabetical order)

Content Hosting	Not all content needs to be hosted by third parties. For example, if a website uses a third-party font library, the site could be changed to use either default browser fonts or self-hosted fonts to provide an immediate privacy gain for visitors. It may not be feasible to self-host all content. Video hosting, for example, can be very resource intensive and expensive. However, website owners may wish to investigate which third-party video hosts are out there, and how (if at all) they use visitor data. YouTube, for example, offers "Privacy Enhancing Mode," which "allows you to embed YouTube videos without using cookies that track viewing behavior."[24] However, it is disabled by default and so action is required by the website's publisher to benefit from this privacy gain. **Ease of Migration: EASY**
Design Optimization	Self-hosted design testing tools may not offer all of the functions of those of third parties. It might be a better idea to review the privacy policy of any tools that are used, and to make sure they respect the privacy of visitors. As most of these services are paid tools and involve some kind of contractual relationship, it is likely possible to find a design testing tool that does not share or unnecessarily retain audience data. **Ease of Migration: Possibly Not Necessary**
Recommendation Systems	Recommendation engines are a major source of third-party cookies. While they may look harmless and may improve the appearance and functionality of a website, they can also be Trojan horses, inserting hundreds of third-party trackers into every page of a website (including malware, in some instances). These tools can also be used to direct viewers to content that belongs to another publisher, or to inject misinformation or misleading stories onto your webpage. Website owners who cannot develop their own content recommendation engines may be better off eliminating the use of these third-party tools altogether. **Ease of Migration: DIFFICULT**
Social Media Share Buttons	Some of the most significant privacy violations we saw came from deploying social media share buttons. Embedding a Facebook "like" button into a webpage enables that platform to be able to link a user's reading and browsing history to their Facebook account. However, it is not necessary to use Facebook's default "like" button to encourage sharing news articles on Facebook. There are self-hosted social sharing plug-ins that transmit less personal information to third parties that may be used instead. **Ease of Migration: EASY**
Visualizations	Many third-party visualization tools offer self-hosted deployments, though some technical knowledge may be required to synchronize their visual interfaces with local deployment. Others do not pose significant privacy risks because as paid tools they offer contractual assurances around how they will or will not use data. **Ease of Migration: EASY**

Conclusion and Recommendations

A year after a new wave of data protection regulations such as the European Union's GDPR have come into effect, the websites of media outlets continue collecting great volumes of personal information—but often unintentionally, and typically for other parties. That so many media institutions have failed to safeguard this asset—to both protect the privacy and safety of their readers and to be in a better negotiating position with advertisers—suggests that education, capacity building, and direct support of independent news outlets is needed to improve their analytics activities and ensure that they safeguard reader privacy.

While it will take additional effort by website operators, safeguarding the privacy rights of visitors may be good for business.

Indeed, the findings in this report reveal that the current level of preparedness among smaller media companies in the Global South to protect their readers from being identified and to protect the commercial value of their analytics data is low. Experts consulted for this report said that this is unfortunate as many of the most common data analytic practices that independent media outlets engage in and benefit from do not require the use of personal information at all.

While it will take additional effort by website operators, safeguarding the privacy rights of visitors may be good for business. This, in turn, could help improve the commercial viability of independent media. Research from the World Economic Forum shows that in the $3 trillion global data economy, online news sites are unusually well-trusted relative to search engines, social networking platforms, and even financial institutions.[25] Yet, so far, a handful of players like Facebook and Google have led the business of online data collection and marketing, which has allowed them to harness the value of the data that they have collected from websites whose content they do not control. This research reveals that independent media websites commonly leak personal information to third parties under the following scenarios: when analytic tools are used to measure audience demographics, when split tests are conducted to experiment with new website design features, when social media "share" buttons are embedded into webpages, or when content recommendation engines are deployed to personalize a website's content to make it relevant to the reader.

Maintaining and building upon the high degree of trust that exists between a publisher and its readers may be the long-term solution

to addressing data privacy concerns while simultaneously developing an effective business model. "Media sustainability in the long run is actually going to be based on your own user base rather than that of advertisers," said Tanja Maksic, a researcher with the Balkan Investigative Reporting Network.[26] This, in turn, could result in digital publishers developing greater negotiating power and leverage with third-party advertising exchanges or, alternatively, being able to sell premium subscriptions to readers. "Look at your internal organization, what you are collecting, what you are doing with it, how you are protecting it," said Maksic, "and craft your economic sustainability around your user base and meet their demands and their standards." Louise Marie Hurel, an internet governance researcher at Brazil's Igarapé Institute, agreed. "I think this is inevitable really. Enforcement of data protection laws may be ad hoc, but that same degree of uncertainty should not mark your relationship with users who are accessing your content."[27]

From a media development perspective, news organizations around the world need better support and training on how to safeguard their valuable audience data, both to protect the privacy rights of their readers and for commercial purposes. Practical primers on complying with data protection laws and regulations, sharing best practices, regularly auditing websites to understand what tracking devices are present and what they are doing with data, and developing and exchanging benchmarking information could all help smaller media outlets find a competitive advantage over the advertising networks and platforms that have absorbed the bulk of online advertising dollars to date. Ultimately, a new business model for journalism will require both savvy use of data and an abiding respect for readers' privacy.

From a media development perspective, news organizations around the world need better support and training on how to safeguard their valuable audience data, both to protect the privacy rights of their readers and for commercial purposes.

Appendix A—Tracking Domains Identified Through Study

The following 167 companies were found to be sending data to 246 domain names by way of third-party cookies installed on the websites of independent media outlets.

Tracking Domain	Tracking Company
33across.com	33Across
undertone.com	33Across
securedvisit.com	4Cite Marketing
4finance.com	4finance
acuityplatform.com	Acuity
pippio.com	Acxiom
addthis.com	AddThis
addtoany.com	AddToAny
ipredictive.com	Adelphic
adform.net	Adform
adgrx.com	AdGear
adblade.com	Adiant
adition.com	Adition
adkernel.com	AdKernel
admedo.com	Admedo
admixer.net	Admixer
everesttech.net	Adobe
demdex.net	Adobe
tubemogul.com	Adobe
adotmob.com	Adot
adriver.ru	AdRiver
akamaized.net	Akamai Technologies
amazon-adsystem.com	Amazon
turn.com	Amobee
adnxs.com	AppNexus
metadsp.co.uk	Avid Media
avocet.io	Avocet
widespace.com	Azerion
bettingpartners.com	Best Partners
betweendigital.com	Between Digital
bidswitch.net	BidSwitch
bttrack.com	Bidtellect
bidtheatre.com	BidTheatre
ml314.com	Bombora
brightcove.net	Brightcove
e-planning.net	Caraytech
cardlytics.com	Cardlytics
casalemedia.com	Casalemedia
sitescout.com	Centro
clevernt.com	Clever Advertising
clickagy.com	Clickagy
onaudience.com	Cloud Technologies
cloudflare.com	Cloudflare
cogocast.net	Cogo Labs
colpirio.com	Colpirio
fwmrm.net	Comcast
scorecardresearch.com	Comscore
connexity.net	Connexity
dotomi.com	Conversant
crazyegg.com	Crazy Egg
ctnsnet.com	Crimtan
criteo.com	Criteo
cxense.com	Cxense
videmob.com	Cydersoft
dable.io	Dable
w55c.net	dataxu
dochase.com	Dochase
adsymptotic.com	Drawbridge
media6degrees.com	Dstillery
dtscout.com	DTS

Tracking Domain	Tracking Company
sociomantic.com	dunnhumby
dyntrk.com	Dynadmic
eboundservices.com	eBound
emxdgt.com	EMX
esquemas.com	Esquemas
exelator.com	eXelate
tribalfusion.com	Exponential
exposebox.com	ExposeBox
eyeota.net	Eyeota
eyereturn.com	Eyereturn Marketing
eyeviewads.com	Eyeview
facebook.com	Facebook
atdmt.com	Facebook
lijit.com	Federated Media Publishing
fidelity-media.com	Fidelity Media
stickyadstv.com	FreeWheel
gemius.pl	Gemius
genieesspv.jp	Geniee
adhigh.net	Getintent
getsocial.io	GetSocial
doubleclick.net	Google
google.com	Google
youtube.com	Google
gstatic.com	Google
groovinads.com	GroovinAds
gumgum.com	GumGum
histats.com	Histats
digitru.st	IAB
ibillboard.com	iBILLBOARD
id5-sync.com	ID5
netmng.com	IgnitionOne
360yield.com	Improve Digital
impdesk.com	Infectious Media
innity.com	Innity
innovid.com	Innovid
inskinad.com	Inskin
insticator.com	Insticator
onthe.io	IO Technologies
ispot.tv	iSpot
izooto.com	iZooto
justpremium.com	JustPremium
daum.net	Kakao
creative-serving.com	KPN
latinongroup.com	LatinOn
loopme.me	LoopMe
crwdcntrl.net	Lotame
lytics.io	Lytics
list-manage.com	MailChimp
mookie1.com	Media Innovation Group
media.net	Media.net
mathtag.com	MediaMath
marketgid.com	MGID
bing.com	Microsoft
linkedin.com	Microsoft
mixmarket.biz	MixMarket
narrative.io	Narrative
navdmp.com	Navegg
agkn.com	Neustar
unrulymedia.com	News Corporation
toast.com	NHN

Tracking Domain	Tracking Company
onesignal.com	OneSignal
bluekai.com	Oracle
zemanta.com	Outbrain
owneriq.net	ownerIQ
paypalobjects.com	PayPal
paystack.com	Paystack
paystack.co	Paystack
adrta.com	Pixalate
playground.xyz	Playground XYZ
powerlinks.com	PowerLinks
infogram.com	Prezi
pubmatic.com	PubMatic
contextweb.com	PulsePoint
quantserve.com	Quantcast
metype.com	Quintype
quora.com	Quora
po.st	R1Demand
gwallet.com	RadiumOne
rating-widget.com	RatingWidget
rawgit.com	RawGit
recreativ.ru	Recreativ
reddit.com	Reddit
reembed.com	reEmbed
republer.com	Republer
reson8.com	Resonate Networks
responsivevoice.org	ResponsiveVoice
retargetly.com	Retargetly
rubiconproject.com	Rubicon Project
rundsp.com	RUN
scribd.com	Scribd
rutarget.ru	Segmento
sharethis.com	ShareThis
simpli.fi	Simpli.fi
sinoptik.ua	Sinoptik
serving-sys.com	Sizmek
skplanet.com	SK Planet
sonobi.com	Sonobi
spotify.com	Spotify
spotxchange.com	SpotX
stackadapt.com	StackAdapt
sundaysky.com	SundaySky
survata.com	Survata
taboola.com	Taboola
tailtarget.com	Target Audiences and Insights Lab
teads.tv	Teads
tapad.com	Telenor
tickcounter.com	TickCounter
adsrvr.org	Trade Desk
tradingview.com	TradingView
programattik.com	Türk Telekom
twitter.com	Twitter
tynt.com	Tynt
upravel.com	Upravel
advertising.com	Verizon
yahoo.com	Verizon
uplynk.com	Verizon
adtech.de	Verizon
weborama.com	Weborama
weborama.fr	Weborama
pushcrew.com	Wingify

Appendix B—Small Publishers Studied

We analyzed the websites of the following small publishers on April 11, 2019:

Publisher	Country
The Bubble	Argentina
Revista Anfibia	Argentina
La Izquierda Diario	Argentina
El Cronista	Argentina
Página/12	Argentina
Publica	Brazil
Revista Fórum	Brazil
Brasil 247	Brazil
Intercept Brasil	Brazil
O Antagonista	Brazil
Mada Masr	Egypt
Egypt Independent	Egypt
Daily News Egypt	Egypt
Akhbar el-Yom	Egypt
Ahram Online	Egypt
Khabar Lahariya	India
The Caravan	India
Scroll	India
The Wire	India
The Quint	India
Tempo	Indonesia
RUAI TV	Indonesia
Antara News	Indonesia
Tirto	Indonesia
Coconuts Jakarta	Indonesia

Publisher	Country
Thika Town Today	Kenya
Africa Uncensored	Kenya
Hivisasa	Kenya
The Elephant	Kenya
Talk Africa	Kenya
Daily Trust	Nigeria
Premium Times	Nigeria
Sahara Reporters	Nigeria
TheCable	Nigeria
Stears Business	Nigeria
Enab Baladi	Syria
ANA Press	Syria
Shaam News Network	Syria
Aleppo Today	Syria
Ain Al-Madinah	Syria
Ukraine World	Ukraine
Euromaidan	Ukraine
Novoye Vremya	Ukraine
Hromadske Radio	Ukraine
Zerkalo Nedeli	Ukraine
MercoPress	Uruguay
LaRed21	Uruguay
Diario El Telégrafo	Uruguay
Brecha	Uruguay
Búsqueda	Uruguay

Appendix C—Large Publishers Studied

We analyzed the websites of the following larger publishers on April 11, 2019:

Publisher	Country
Infobae	Argentina
El Intransigente	Argentina
Clarín	Argentina
La Nación	Argentina
Perfil	Argentina
Globo	Brazil
Metrópoles	Brazil
Jornal do Brasil	Brazil
UOL	Brazil
Estadão	Brazil
Al Gomhuria	Egypt
Youm7	Egypt
Sada El Balad	Egypt
Al Masry al Youm	Egypt
El Fagr	Egypt
Times of India	India
Manorama Online	India
NDTV	India
Hindustan Times	India
News 18	India
The Jakarta Post	Indonesia
Tribun News	Indonesia
Detik	Indonesia
Kompas	Indonesia
Liputan 6	Indonesia
Daily Nation	Kenya

Publisher	Country
Business Daily	Kenya
Capital FM	Kenya
Kenya Broadcasting Corporation	Kenya
Kenya News	Kenya
Punch Newspaper	Nigeria
Vanguard News	Nigeria
The Guardian	Nigeria
This Day	Nigeria
P.M. News	Nigeria
SANA	Syria
Syria Report	Syria
Raialyoum	Syria
Zamanalwsl	Syria
Aks Alser	Syria
Ukrainian Independent Information Agency	Ukraine
Ukrainian News	Ukraine
Segodnya	Ukraine
Interfax-Ukraine News Agency	Ukraine
Kyiv Post	Ukraine
El País	Uruguay
El Observador	Uruguay
OK Diario	Uruguay
Portal 180	Uruguay
Ecos Diarios Necochea	Uruguay

Appendix D—Study Setup and Testing Parameters

Setup

Tests using OpenWPM were initially conducted on a sample of 40 websites on March 24, 2019, and repeated on April 11, 2019, using the full group of 100 websites. OpenWPM was installed from GitHub using Git revision b3ead7e38892095950806e8bcbb2e1129c27ca96.

Tests were performed using the Kubuntu 18.04 operating system, with Python 2.7.15rc1 and Python 3.6.7 and Firefox 67.0b4. Testing was done under VPN connection.

Testing Parameters

The OpenWPM "demo.py" script was used as a template and modified. The value of NUM_BROWSER was set to 1 to use only one web browser and to be sure that the websites were crawled in the given order. The browser was not headless. Flash was enabled. Cookie_instrument (experimental) was enabled.

The following is the sequence of commands used for each website:

- Visit the homepage and wait for 120 seconds
- Dump flash cookies
- Dump profile cookies

To prevent data contamination, the previously generated SQLite database was deleted before each new recording.

Endnotes

[1] Alan Westin, *Privacy and Freedom* (1967), 7.

[2] Nonpersonal data include information that does not concern a natural person. For instance, a data set of daily temperatures in a city would constitute nonpersonal data.

[3] Anonymized data include information that was originally personal information but has been transformed in such a way that the link between it and the natural person has been cut. Many data analysis activities are performed on anonymized data.

[4] European Union Law, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," Article 4 (1), April 27, 2016, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex:32016R0679.

[5] This is known as the Mosaic Effect. In a study published in *Science* in 2015, researchers found that four data points are enough to uniquely and accurately reidentify an individual in 90 percent of cases. See Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh, and Alex "Sandy" Pentland, "Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata," *Science* 347, no. 6221 (January 2015): 536-539, http://science.sciencemag.org/content/347/6221/536.full.

[6] The number of countries with adopted data protection laws as of March 27, 2019, is 134. The geographical distribution of the 134 laws is the following: Africa (26), Asia-Pacific (22), Europe (54), Latin America and Caribbean (23), Middle East (7), and North America (2).

[7] In alphabetical order: Barbados, Belarus, Brunei, Dominica, Ecuador, Egypt, El Salvador, Ethiopia, Falkland Islands, Grenada, Guatemala, Honduras, Indonesia (substantial revision to existing law), Jamaica, Jordan, Kenya, Montserrat, Nigeria, Saint Helena, Saint Kitts and Nevis, Saudi Arabia, Swaziland, Tanzania, Virgin Islands, Zambia, and Zimbabwe.

[8] Mark Scott and Laurens Cerulus, "Europe's New Data Protection Rules Export Privacy Standards Worldwide," *Politico*, January 31, 2018, https://www.politico.eu/article/europe-data-protection-privacy-standards-gdpr-general-protection-data-regulation/.

[9] Robert Gellman, *Fair Information Practices: A Basic History* (April 10, 2017), available at SSRN: https://ssrn.com/abstract=2415020.

[10] Secretary's Advisory Committee on Automated Personal Data Systems, *Records, Computers and the Rights of Citizens,* Office of the Assistant Secretary for Planning and Evaluation, US Department of Health and Human Services, July 1, 1973, https://aspe.hhs.gov/report/records-computers-and-rights-citizens.

[11] Sarah Gordon and Aliya Ram, "Information Wars: How Europe Became the World's Data Police," *Financial Times*, May 20, 2018, https://www.ft.com/content/1aa9b0fa-5786-11e8-bdb7-f6677d2e1ce8.

[12] Rebecca Hill, "Washington Post Offers Invalid Cookie Consent under EU Rules – ICO," *The Register,* November 19, 2018, https://www.theregister.co.uk/2018/11/19/ico_washington_post/.

[13] S. Rana interviewed by A. Férdeline via Skype, March 14, 2019.

[14] Julie Brill, "Microsoft's Commitment to GDPR, Privacy and Putting Customers in Control of Their Own Data," Microsoft, blog post, May 21, 2018, https://blogs.microsoft.com/on-the-issues/2018/05/21/microsofts-commitment-to-gdpr-privacy-and-putting-customers-in-control-of-their-own-data/.

[15] General Data Protection Regulation, "Processing of Personal Data Solely for Journalistic Purposes or for the Purposes of Academic, Artistic or Literary Expression," Recital 153, via Intersoft Consulting, https://gdpr-info.eu/recitals/no-153/.

[16] I. Avadani interviewed by A. Férdeline via Skype, March 19, 2019.

[17] "OCCRP Strongly Objects to Romania's Misuse of GDPR to Muzzle Media," Organized Crime and Corruption Reporting Project, November 9, 2018, https://www.occrp.org/en/40-press-releases/presss-releases/8875-occrp-strongly-objects-to-romania-s-misuse-of-gdpr-to-muzzle-media.

[18] General Data Protection Regulation, Recital 153.

[19] A. Alzghoul interviewed by A. Férdeline via Skype, March 14, 2019.

[20] Please note that this is an oversimplification of the process. Please refer to the following technical document for a more complete explanation of how HTTP requests work: R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, "Hypertext Transfer Protocol," IETF, June 1999, https://www.ietf.org/rfc/rfc2616.txt.

[21] See, for instance, P. Laperdrix,W. Rudametkin, and B. Baudry, "Beauty and the Beast: Diverting Modern Web Browsers to Build Unique Browser Fingerprints," 37th IEEE Symposium on Security and Privacy, 2016, https://www.ieee-security.org/TC/SP2016; N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna, "Cookieless Monster: Exploring the Ecosystem of Web-Based Device Fingerprinting," 34th IEEE Symposium on Security and Privacy, 2013, https://ieeexplore.ieee.org/document/6547132/.

[22] For further information on the intricacies of how the OpenWPM tool works, please see the Princeton Web Census website at https://webtransparency.cs.princeton.edu/webcensus/.

[23] V. Pavel interviewed by A. Férdeline via email, February 26, 2019.

[24] "Turn on Privacy-Enhanced Mode," YouTube Help, 2019, https://support.google.com/youtube/answer/171780?visit_id=0-636595692661723869-3019304114&rd=1.

[25] Vasudha Thirani and Arvind Gupta, "The Value of Data," World Economic Forum, 2017, https://www.weforum.org/agenda/2017/09/the-value-of-data/.; In a 2014 poll, 56 percent of respondents in countries with less than 25 percent internet penetration answered 5, 6, or 7 on a seven-point trust scale to indicate that they trust online news sites to protect their personal data. This compares with 40 percent trusting search engine companies, 37 percent trusting companies that provide social networking services, and 29 percent trusting online marketers and advertisers. The only stakeholder group more trusted than the media were banks and financial institutions, who were trusted by 61 percent of respondents. Source: William H. Dutton, Ginette Law, Gillian Bolsover, and Soumitra Dutta, *The Internet Trust Bubble: Global Values, Beliefs and Practices* (World Economic Forum, 2014), http://www3.weforum.org/docs/WEF_InternetTrustBubble_Report2_2014.pdf.

[26] T. Maksic interviewed by A. Férdeline via Skype, February 8, 2019.

[27] L. Hurel interviewed by A. Férdeline in Kobe, Japan, March 13, 2019.

Center for International Media Assistance

NATIONAL ENDOWMENT FOR DEMOCRACY
1025 F STREET, N.W., 8TH FLOOR
WASHINGTON, DC 20004

PHONE: (202) 378-9700
EMAIL: CIMA@ned.org
URL: https://cima.ned.org

CIMA
CENTER FOR INTERNATIONAL MEDIA ASSISTANCE

NED
National Endowment for Democracy
Supporting freedom around the world