



Third-Party Web Content on EU News Sites: Potential Challenges and Paths to Privacy Improvement

Authors: Timothy Libert and Rasmus Kleis Nielsen

Introduction

Most of the content, products, and services we use online, whether provided by media organisations or by technology companies, are funded in whole or in part by advertising and various forms of marketing. Such advertising, and more broadly many other online services, increasingly involves the collection and analysis of large amounts of personal information that we wittingly or unwittingly let websites collect as we use them, personal information that is in turn often shared with various third parties through third-party elements and cookies included when we load a website, oftentimes without users' consent or even disclosure.

This factsheet provides an overview of third-party web content practices involved in such data collection. We focus on a sample of major news sites in seven EU countries, compare news sites with popular sites, and explore differences between the countries covered. Analysis of millions of content requests and cookies helps us measure the amount of third-party elements and cookies included on news sites and other popular sites, identify the companies involved, and categorise the different types of third-party content found. Across all countries, we find that news sites include significantly higher volumes of third-party content and cookies compared with other popular sites. On average, news sites load four times as many third-party domains (40 vs 10) and set almost eight times as many third-party cookies (81 vs 12).

Advertisers argue that the collection of data through

practices like the ones we analyse here allows for more tailored and effective advertising, while technology companies say it enables innovation and the provision of personalised services that users value. Media organisations are beginning to gather and use similar data to inform editorial decision-making, help with product development, and meet advertisers' demand for data-based targeting. Increasingly, however, some users and policymakers are questioning the scale and scope of data collection online, the extent to which personal information is shared with third parties, and the ways in which data is used. These concerns are reflected in the European Union's General Data Protection Regulation (GDPR), which will come into force on 25 May 2018.

Much of the discussion around data collection and data sharing with third parties has understandably focused on large technology companies like Facebook and Google, who collect data globally on billions of users. But, as we document here, data is collected across the web by many different sites, including major news sites, and is often shared with various third parties in the process.

Based on analysis of 500 popular sites and prominent news sites in each of the seven countries covered (Finland, France, Germany, Italy, Poland, Spain, and the UK), we show that:

- Third-party content is present on the vast majority of websites and is especially prevalent on news sites. Over 95% of news sites across the

seven countries contain third-party content, and over 90% set at least one third-party cookie. In every country, the numbers for our sample of major news sites are consistently higher than those for the 500 most popular sites in the same country.

- News sites expose users to many third-party domains concurrently, ranging from an average of 24 third-party domains per news site in Italy to an average of 50 on UK news sites. Likewise, news sites have many third-party cookies on each page: ranging from an average of 42 per page in Italy to 90 per page in the UK. Again, the numbers for our sample of major news sites are consistently higher than those for the 500 most popular sites in each country.
- News sites based on different funding models include different numbers of third-party domains and cookies. A strategic sample of news sites in the UK and Germany reveals that popular newspapers (often heavily reliant on advertising revenues) have the most third-party content and cookies, upmarket newspapers (that rely on advertising as well as subscription revenues) have a medium amount, and public service media (that are not reliant on advertising or subscriptions) have relatively few.
- A limited number of US-based technology companies have third-party content on a significant percentage of all pages analysed (including both news sites and popular sites). The two most prominent are Google (on 87% of all pages) and Facebook (on 40%), far ahead of Amazon (17%) and Twitter (15%) in third and fourth place. Across all sites, the top ten companies hosting third-party content are based in the United States. The highest-ranked EU-based company, Criteo, is at position 13 and has content on 7% of sites.
- Third-party content comes in several prominent varieties: advertising and marketing, audience measurement, design optimisation, social media, content recommendation, and content hosting (including JavaScript, font, image, and video hosting), most of them more widely (and often more intensely) used by news sites than other sites.
- Different types of third-party content place cookies at different rates and represent different levels of risk to user privacy. We identify a range of possible data protection and privacy issues facing news sites and categorise them in terms of how much effort it will require to address them.

Common Data Protection Frameworks for Online Privacy

As the collection and analysis of growing amounts of data and personal information are becoming increasingly integral to most of the services we use and the advertising that we see when we use digital media, discussions around the norms, risks, and proper regulatory framework have intensified.

Privacy norms and risks have been understood for some time and numerous regulations exist to protect individuals. In 1973, the US Department of Health, Education, and Welfare published a report detailing five 'Fair Information Practice Principles' (FIPPS), which have proven hugely influential. The principles are (1) notice/awareness, (2) choice/consent, (3) access/participation, (4) integrity/security, and (5) enforcement/redress. Among the guidelines and laws influenced by FIPPS are the OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, the Asia-Pacific Economic Cooperation forum guidelines, the 1995 EU Data Protection Directive, and the EU GDPR, which will cover 28 member states with a combined population of more than half a billion. Likewise, self-regulatory guidelines promulgated by online marketers reflect FIPPS concepts even if some critics argue they fail to embrace the spirit of the principles.

The European Union's GDPR, adopted in 2016 and coming into force in May 2018 (replacing the older 1995 Data Protection Directive), is a new law, but it may be viewed as part of the evolution of well-established approaches to privacy regulation. As with earlier privacy frameworks, the GDPR specifies a range of restrictions on how the data of 'natural persons' may be collected, stored, and processed online. The GDPR likewise places substantive limits on how various categories of data may be processed and requires affirmative 'opt-in' from users in many cases.

The most notable aspect of the new regulation may be that it allows for penalties of up to €20 million or 4% of global revenue (whichever is higher) for companies found to be violating it. The GDPR is, therefore, being taken very seriously by all businesses, European or non-European, which operate in the EU. As one reporter has noted, 'the likes of Google, Facebook, and the world's largest financial institutions [have] spent years investing in new compliance structures and hiring hundreds of new lawyers, coders, and designers to make sure they would follow the letter of the upcoming law'.¹

¹ <https://www.politico.eu/article/gdpr-rules-europe-facebook-data-protection-privacy-general-data-protection-regulation-cambridge-analytica/>

But because data collection and analysis is increasingly integral to most of the content, products, and services we use online, the GDPR will not only impact large technology companies and various information brokers. It also puts the spotlight on the data-collection, data-sharing, and advertising practices of many media organisations, including news sites (MacGregor and Zylberberg, 2018).

To explore these practices here we begin with a quick primer on the basic technical features of web pages (some readers may want to skip this section). We then identify the most important kinds of third-party content that feature on many news sites and present the main findings from our analysis.

How Web Pages Work: First- and Third-Party Content

Data-collection and online privacy concerns are rooted in the architecture of the web. When users access a web page they are not downloading a finished media object, the way one might pick up a printed newspaper or download a PDF file. Rather, they are accessing a set of instructions for how to build the page written in the Hypertext Markup Language (HTML). A user's web browser reads the HTML instructions, downloads any additional content needed for the page, such as images and JavaScript code, and then displays the page.

Web content is delivered by either a first or third party. First-party content comes from the same location as the HTML file itself, and this is the address a user sees in their browser window. For example, a hypothetical news site at the address 'http://example.com' may feature an image which is downloaded from 'http://example.com/newsimage.jpeg'.

In contrast, third-party content is downloaded from a different internet address, and in many cases, a different company, than the website a user is visiting. If 'http://example.com' includes a video which is hosted at the address 'http://video-hosting.com/newsvideo.mp4', this means 'video-hosting.com' is a third-party content host, and 'newsvideo.mp4' is third-party content. It is still possible that the first party originally uploaded the content, but a user gets it from the third party.

One of the most powerful features of HTML is that it allows for first- and third-party content to be downloaded automatically by a browser and combined to make a web page in milliseconds.

When the web browser downloads content it sends an 'HTTP request' to either a first- or third-party server. The server is then able to determine a user's IP address as well as information on the type of computer, operating system, and the URL of the page being viewed. In the context of a first-party request, the user likely has an implicit understanding they are disclosing data to the website being visited. It is less clear that downloading third-party content – which the user has not directly initiated – transmits the address of the page being viewed to parties that the user is likely not aware of. This process is at the core of how web tracking works.



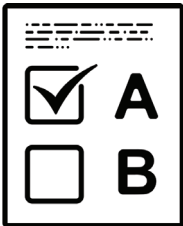

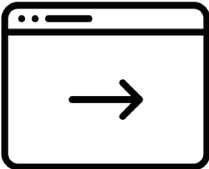
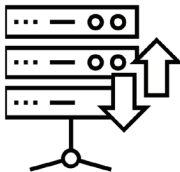
Beyond HTTP request data, it is possible for third parties to enhance the quality of their tracking by leveraging identifiers such as cookies, data set by a content host in the user's web browser to create a unique identifier which can be used to track the user on the web. Likewise, so-called 'browser fingerprints' may be used to track users on the web. It is important to note that the data generated from a single request do not exist in isolation and may represent a single point in a much larger constellation of data about a given user.

Third-Party Web Content: A Categorisation of Purpose and Use

Websites rely on third-party content for a number of uses ranging from knowing how many people visited a given page to making sure the correct font is used in an article's headline. To gain deeper insights into how the third-party content ecosystem works, and what this means for privacy under the GDPR, we examine several categories of third-party web content.

It is essential to note at the outset that the categorisations used in this study reflect the immediate purpose of specific third-party content from the perspective of website operators and visitors. In other words, these categories answer the question of why the owner of the site chose to include the third-party content and what it does for users. The answer to this question may be different from the ultimate reason a third party developed the content. For example, a website developer may use a social media 'share' button to help users share content, but a social network may use the button to track user behaviour and target advertisements.

Six dominant use categories are explored in this study: advertising and marketing, audience measurement, design optimisation, social media,

	<p>Advertising and Marketing: Most websites carry advertising, and some are paid to include third-party marketing content so that marketing firms may monitor a specific individual's interests, locations, purchasing habits, and lifestyle choices. For many websites, especially online news sites, advertising and other income from third-party marketing content is an essential source of revenue. Programmatic advertising exchanges such as Google's DoubleClick and AdSense, as well as platforms such as AppNexus, the Nielsen subsidiary VisualDNA, and Acxiom's LiveRamp, offer a range of different advertising and marketed related-services.</p>
	<p>Audience Measurement: This category of content is often composed of invisible image files, called tracking pixels, and JavaScript code which allows third parties to determine the size and composition of the audience visiting a given site. This may include aggregate statistics as well as detailed information on the demographic characteristics of visitors, their geographic locations, and the actions they take while visiting a site. Widely used audience measurement tools include Google Analytics, Adobe Analytics Cloud, and Quantcast.</p>
	<p>Design Optimisation: Many website developers seek to improve their sites on an ongoing basis so that users are able to find the information they need, navigate more effectively, and load pages faster. Design optimisation tools give developers insights into how their site is performing with actual users and allow them to conduct tests to determine the impact of planned design changes (often called 'A/B Testing'). These tools, from companies such as New Relic and Optimizely, often load third-party JavaScript on a page to conduct advanced analysis of user behaviour on behalf of the website developer.</p>
	<p>Social Media: Many websites contain social media 'widgets', which are blocks of code that facilitate the sharing of the page's URL and content, as well as embedding social media within the page. These widgets are primarily included in the page by downloading icons and JavaScript code from third parties. The most widely used social media sharing widget comes from the biggest social media service: Facebook. Others are provided by companies such as AddThis, which provide widgets which allow users to connect to a wide range of different social networks.</p>
	<p>Content Recommendation: Found primarily on news media websites, content recommendation tools are often composed of boxes sitting below articles which feature links to related articles on the same website, articles on partner sites, as well as links to advertising content. The article links in these widgets are dynamically updated by third-party JavaScript code. The two main content recommendation platforms are Taboola and Outbrain.</p>
	<p>Content Hosting: Content hosting companies such as Amazon Web Services operate data centres – vast, secured buildings where thousands of servers are physically located – which host the websites of content providers such as media organisations. In addition to general-purpose content hosts, there are a range of specialised content hosting services often used to deliver common JavaScript code libraries, images, fonts, as well as videos. Websites may use specialised content services due to perceived ease of use for developers. Content such as video may be particularly difficult to host on a first-party basis due to extremely high volumes of data.</p>

content recommendation, and content hosting. They are described in the boxes above:

As noted, all requests for third-party content on a given web page expose user data and IP addresses, and many of them are central to increasingly important forms of data collection and data analysis, some of which are

increasingly under scrutiny from a privacy perspective. To shed light on these practices, the next section details a means to measure third-party requests, presents extensive findings derived from over 200,000 page loads, and discusses the findings in light of the impending introduction of the GDPR.

Website Selection and Third-Party Content Measurement

Seven countries were selected for this study to represent a mix of population sizes and media markets in the EU. The countries studied are Finland, France, Germany, Italy, Poland, Spain, and the United Kingdom. In each country, we included prominent news sites, selected on the basis of prior work measuring their reach and significance (Newman et al., 2017). The Alexa Web Information Service was used to select the top 500 most popular sites in each country to provide a baseline for comparison.

To measure the presence and nature of third-party content on the selected websites, we used the open-source software tool webXray.² This analyses a page by opening it in the Chrome web browser and creating a new user profile which has no cookies or history. The software then loads the page in Chrome and keeps the page open for 30 seconds, during which time all requests for third-party content are monitored. At no point is the browser interacted with in any way, and no cookie or tracking consent buttons are clicked. Finally, webXray extracts all cookies from the internal Chrome database, records them, and closes the browser.

In the three-month period starting in January 2018, homepages from 3,500 popular sites and over 200 news sites were loaded with the webXray software platform. A total of 203,871 page loads, over 20 million content requests, and 4.5 million cookies were captured and analysed. This time period was chosen to give a baseline measure of the state of third-party web content immediately *prior* to the introduction of the GDPR so that the steps needed to comply with the GDPR may be considered. We plan to repeat the analysis at a later stage to compare sites before and after the GDPR comes into force.

For more details on the sample and methods, please see the appendix.

Top-Level Findings

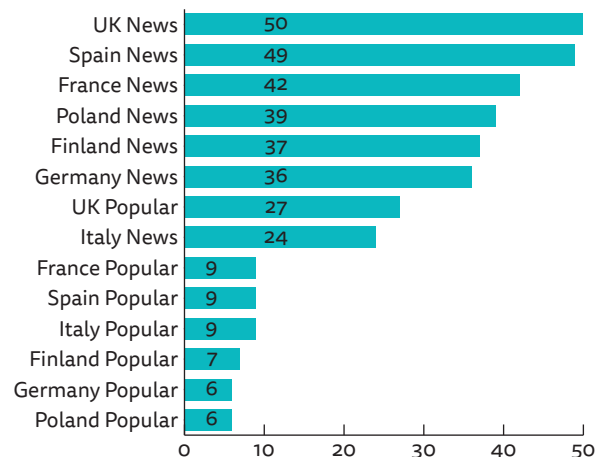
The majority of sites in our sample have third-party content hosted by a number of companies providing a range of services, often related to data collection, data sharing, and data analysis that may impact user privacy. Across all measures, news sites expose users to more third-party and potentially privacy-

compromising content. The analysis thus identifies a range of issues in light of the impending GDPR, as well as a number of ways in which user privacy can be improved.

The vast majority of both popular and news sites in all countries contain third-party content. The lowest percentage of sites with any third-party content is on popular sites in Germany, with 84%.

The percentage of pages with *any* third-party content obscures the *amount* of third-party content on a per-page basis, which is strikingly different when we compare news sites with our baseline of popular sites. Figure 1 illustrates the average number of distinct third-party domains found on each type of page. At the high end, UK news sites load over eight times more third-party content per page than popular sites in Poland at the low end.

Figure 1. Third-party domains per page

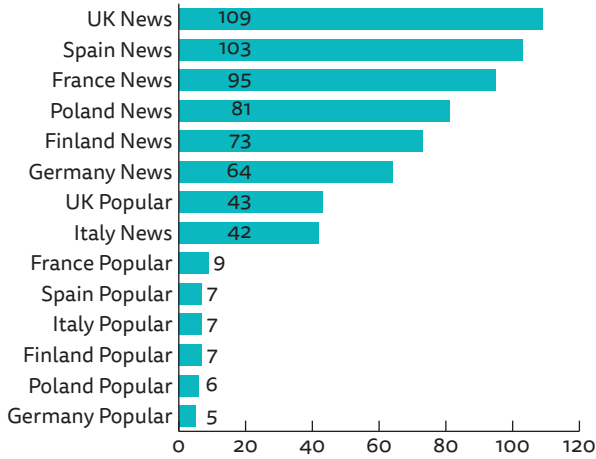


While all requests for third-party content reveal users' IP addresses and something about their browsing interests, this data may not provide high-quality tracking data. To improve the quality of the data collected and increase their utility for both advertising placement and other purposes, third-party cookies may be used to track users between many sites.

Figure 2 shows that the percentage of sites with at least one third-party cookie varies widely, from a low of 45% of Polish sites to over 90% of news sites in all countries. As with third-party content, the number of cookies on a per-page basis varies. Popular sites in Germany have an average of five third-party cookies per page compared with an average of 90 per page on UK news sites, an 18-fold difference.

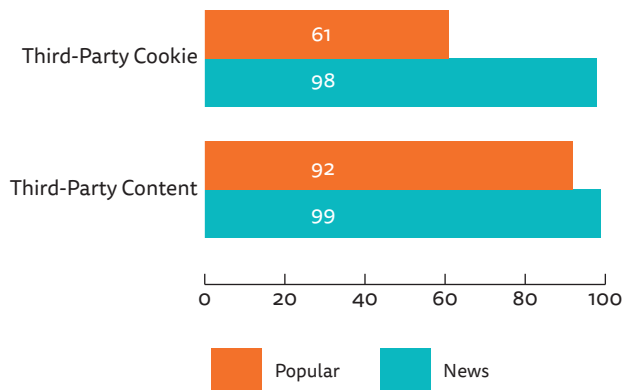
² Those who wish to know more about webXray may visit the project website (<https://webxray.org>) and download the software (<https://github.com/timlib/>).

Figure 2. Cookies per page



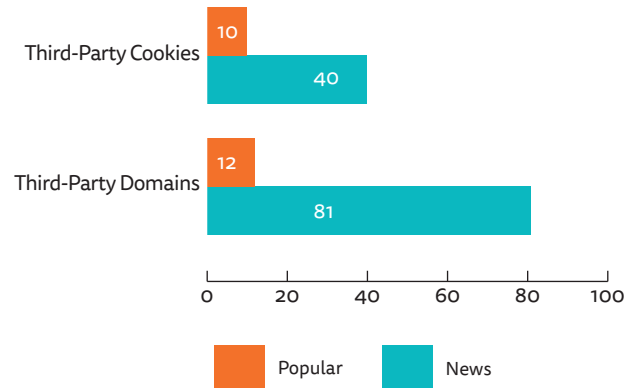
Figures 3 and 4 compare popular sites and news sites overall by averaging each group together across all seven countries. As shown in figure 3, 99% of news sites have at least one third-party request compared with 92% for popular sites, which is fairly similar. Cookies exhibit a much starker difference: 98% of news sites have at least one third-party cookie compared with 61% for popular sites.

Figure 3. News vs popular sites, percentages



The difference between news and other popular sites is most profound when looking at the volume of third-party content requests and cookies per page, captured in figure 4. News sites have an average of 40 third-party requests per page compared with 10 for popular sites. Likewise, news sites have an average of 81 cookies per page compared with 12 for popular sites.

Figure 4. News vs popular sites, counts



It is important to recognise that there is considerable variation underneath these headline findings. Different news sites are run by media organisations with different funding models and strategic priorities. To illustrate the variation, we compare a strategic sample of three different types of news sites in Germany and the UK, one a popular newspaper (*Bild* and the *Daily Mirror*), one an upmarket newspaper (*Süddeutsche Zeitung* and the *Times of London*) and one a public service media organisation (ARD/Tagesschau and BBC News). The homepage of each newspaper was loaded with webXray on April 30th to obtain precise measures.

Across all countries, popular newspapers (which are frequently more reliant on advertising revenues) have the most third-party content and cookies, upmarket newspapers (which rely on advertising as well as subscriptions) have a median amount, and public service media (which benefit from public funding and are thus not dependent on advertising or subscriptions) have relatively few. Again, in each category, sites in the UK have more third-party content and cookies than sites in Germany. The German popular newspaper *Bild*, for example, sets fewer cookies than the upmarket UK newspaper *The Times*. At the outer ends of the spectrum, a visitor to the *Daily Mirror* will have 246 third-party cookies set, over 80 times more than a visitor to the ARD/Tagesschau will receive.

Figure 5. Site comparison



Some sites share users data with many more third parties, sometimes without necessarily being fully across who exactly has access to the data. Recently, the sports site GiveMeSport revealed that an internal audit had found that up to 500 different companies were processing its users’ personal data in ways that could be illegal under GDPR. In an interview with DigiDay, Ryan Skeggs, general manager at GiveMeSport in the UK said: “We looked through all the partners plugged into the site and recognized 8 percent of them — that’s pretty shocking. We didn’t have any idea who 92 percent [of those businesses] were.”³

Prevalence of Specific Companies

The software used for this study, webXray, is able to identify over 400 different companies and services, 270 of which were detected in this analysis. Given the complexity of analysing companies across all countries separately, a single scan of the homepages of all sites was performed in April 2018.

Several companies have third-party content on a significant percentage of all pages. Table 1 shows the top ten companies hosting third-party content, all of which are based in the United States. The two most prominent are Google, which across a range of services is present on 87% of all pages, and Facebook, which is on 40% of all pages. After these two, the reach of individual companies and services declines, with 88% of companies and services identified present on fewer than 5% of all pages analysed. The highest-ranked EU-based company, Criteo, is outside the table at position 13 and has content on 7% of sites.

Table 1. Percentage of sites with content from a company

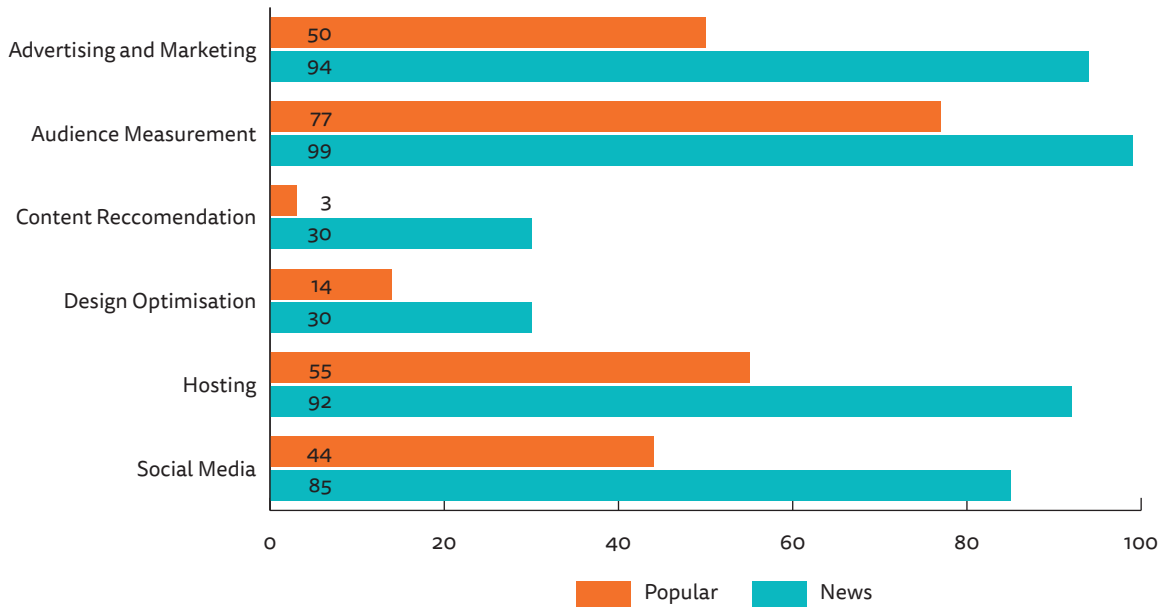
Most prominent companies		
Percentage pages tracked	Owner	Owner country
87	Google	US
40	Facebook	US
17	Amazon	US
15	Twitter	US
13	Oracle	US
12	Adobe Systems	US
12	AppNexus	US
11	Rubicon Project	US
10	Oath	US
10	Cloudflare	US

Google’s widespread presence is partially due to the variety of services and subsidiary companies as shown in table 2. The most prominent service is Google Analytics, on 74% of sites. Google’s advertising network, DoubleClick, is on 47% of sites, and YouTube videos are on 15%.

Table 2. Percentage of sites with content from Google subsidiaries and services

Percentage pages tracked	Service
74	Google Analytics
58	Google APIs
47	DoubleClick
32	Google Tag Manager
26	AdSense
15	YouTube
< 1	Google App Engine
< 1	Blogger

³ <https://digiday.com/uk/publishers-say-theyll-use-gdpr-shed-ad-tech-vendors/>

Figure 6. Content percentages

Types of Third-Party Content

Third-party content is used for a variety of purposes. For this study, six main types of content (advertising and marketing, audience measurement, design optimisation, social media, content recommendation, and content hosting) were identified and examined across news and popular sites in seven EU countries.

In order to simplify the comparison, news and popular sites across the seven countries have been averaged together to reveal trends. Figure 6 shows the percentage of sites with a given type of content. Across all categories, news sites have a higher percentage of pages with that content, indicating deeper reliance on third parties than other popular sites.

The most prominent type of content on all sites is audience measurement, which is on 98% of news and 77% of popular sites. This is driven primarily by the presence of Google Analytics, found on 74% of sites. On popular sites, only two other types of content are on more than half of pages (advertising and marketing, and hosting), whereas on news sites four of the six types are present on between 85 and 98% of all sites. The largest difference is in the content recommendation category, which is highly news specific: 30% of news sites include this type of content compared with only 3% of popular sites.

As with general trends, the percentage of sites having any form of content obscures the number of times such content appears on a *per-page* basis. As figure 7 shows, news sites have more instances of each type

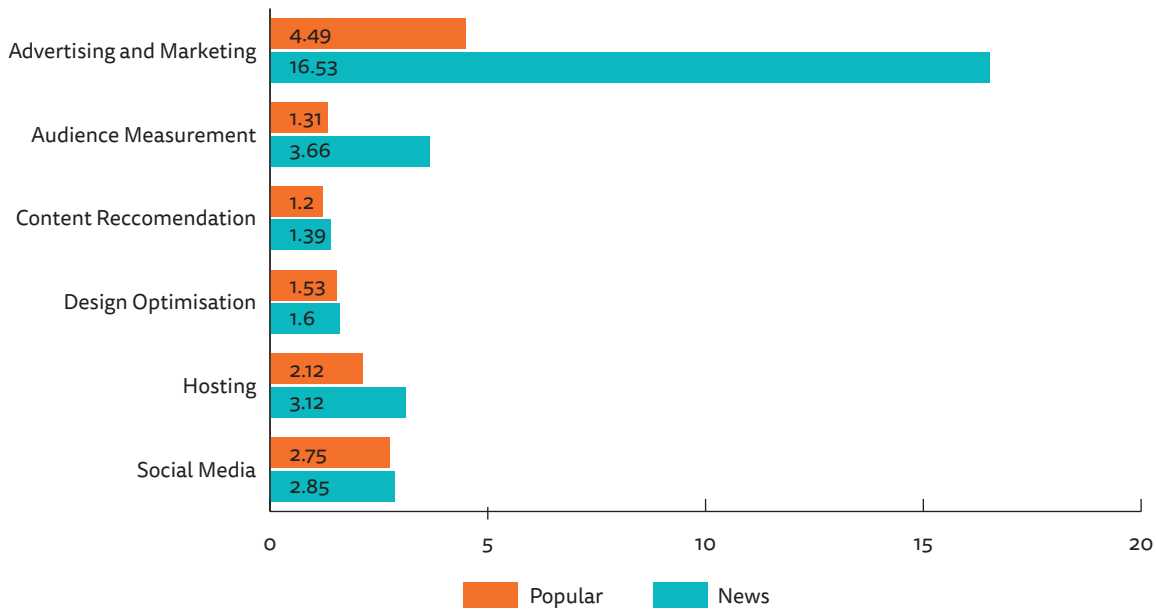
of content. Most striking is that among sites that have some form of advertising and marketing content, news sites have 17 different domains per page compared with 4 on popular sites, a four-fold difference.

Last, across most types of third-party content, news and popular sites have similar rates of cookies being set. The only category with significant variation is audience measurement, in which 45% of news and 13% of popular sites set cookies. This may be a result of news sites configuring audience measurement systems differently, or using systems which do not seek consent for setting cookies.

As noted previously, at no time was consent given for cookies to be set during this analysis. From a data protection and privacy standpoint, setting cookies without explicit and informed user consent is perhaps the most troubling, and widespread, practice detected in this study.

What this Means for the GDPR: Easy Privacy Gains and Confronting Hard Topics

We have produced this factsheet to provide an overview of the third-party web content practices that enable widespread data collection, data sharing, and data analysis across much of the web. Some of these practices will fall under the EU GDPR when it comes into force in May 2018. The precise implications are

Figure 7. Content occurrences

still not clear, in part because of the complexities of the practices covered by the GDPR, in part because the regulation depends on enforcement by national privacy regulators. Different companies and industry associations are advancing different interpretations on what is and is not required under the new regulation and how best to respond. But based on the wording of the GDPR and the practices mapped here, it is possible to indicate some relatively easy steps websites could take to increase users' privacy as well as other areas that will be harder to handle unless explicit user consent is secured.

The GDPR is a large and complicated law, not all of which is pertinent to how third-party web content may be regulated. However, the Recitals of the GDPR, which guide legislative intent, state the following:

- **Recitation 30** establishes that 'Natural persons may be associated with online identifiers provided by their devices, applications, tools and protocols, such as internet protocol addresses, cookie identifiers or other identifiers'.
- **Recitation 39** specifies it 'should be transparent to natural persons that personal data concerning them are collected, used, consulted or otherwise processed'. In particular, the 'identity of the controller and the purposes of the processing' must be made clear.
- **Recitation 40** places restrictions on data processing: 'In order for processing to be lawful, personal data should be processed on the basis

of the consent of the data subject concerned or some other legitimate basis.'

- **Recitation 32** establishes that consent must only occur on an opt-in basis: 'Consent should be given by a clear affirmative act establishing a freely given, specific, informed and unambiguous indication of the data subject's agreement to the processing of personal data relating to him or her, such as by a written statement, including by electronic means, or an oral statement.'

Taken together, these passages indicate that third-party web content may face several obstacles under the GDPR. First, third-party content exposes IP addresses and often results in the transmission of cookie data, potentially putting it within the law's scope. Second, because browsers download third-party content without user interaction, it is often not transparent to users what is happening, raising the need for clear notification whenever data is collected, shared, and stored. Third, for user data to be processed lawfully, consent may be needed. Under GDPR, news sites are what is termed "data controllers" and therefore, according to Article 5(2), "shall be responsible for, and be able to demonstrate compliance with [the principles related to processing of personal data]." Thus, publishers are in the process of reviewing their own practices as well as vetting their ad tech vendors and other partners to ensure all third-party data collection, sharing, and processing is compliant with GDPR.⁴

⁴ <https://digiday.com/uk/publishers-say-theyll-use-gdpr-shed-ad-tech-vendors/>

If no changes are made to any of the websites examined, one potential outcome is the GDPR will result in a barrage of consent notifications for third-party content and data sharing. However, it is entirely possible that content hosted on a first-party basis may not require consent beyond what is needed for the site. One way to reach GDPR compliance may be to migrate content currently hosted by third parties to the first-party website.

Table 3 presents each type of content along with several sub-types of hosted content. Privacy risk for each type of content is judged broadly on a scale of high, medium, to low. This is not a precise quantification and is based on a combination of the rate cookies are set and the possibility that the content may be linked to a user’s identity (as is the case with social media widgets). Likewise, it is important to note that risk itself is often dependent on social class and context: an affluent individual from a dominant social group may not face the same forms of discrimination as those in marginalised groups (Turow, 2011). The high number of cookies found on popular newspapers’ web sites compared to upmarket ones only further underline that different users are differently affected. The second column presents a high, medium, low scale of the difficulty of migrating the content. Again, this is a broad assessment and details will vary from site to site and instance to instance.

Table 3. Types of content, privacy risk, difficulty of converting to first party

Purpose	Privacy risk	Effort to replace
Audience measurement	MEDIUM	MEDIUM
Content recommendation	HIGH	HIGH
Design optimisation	MEDIUM	MEDIUM
Social media	HIGH	LOW
Marketing	HIGH	HIGH
Hosting: Core	LOW	HIGH
Hosting: JavaScript	LOW	LOW
Hosting: Font	LOW	LOW
Hosting: Image	MEDIUM	LOW
Hosting: Video	MEDIUM	HIGH

For sites focused on improving privacy especially in light of the GDPR, content which ranks ‘low’ on the effort scale could be prioritised for migration. Hosted JavaScript files, fonts, and images all have low to medium privacy risk, and in some cases changing a

single line of code may provide immediate privacy gains. Similarly, social media buttons frequently set cookies and may link browsing data directly to users’ profiles, representing a high privacy risk. While social media companies provide code to enable sharing, it is possible to implement widgets on a first-party basis which facilitate social sharing. Even if social media companies would prefer sharing to happen with their widgets, they have no interest in preventing sharing.

Two types of content, audience analytics and design optimisation, we place in the ‘medium’ category. Both of these types of content rely on complex back-end systems to process data. Doing this in-house will require more effort and may not deliver the same functionality. Similarly, there are few mature options for open-source design optimisation, and doing this without reliance on third parties will require more effort. Migrating content in the ‘medium’ effort range is non-trivial. Effort may be measured in months and will often also require significant investment.

Finally, content in the ‘high’ category may be nearly impossible to host on a first-party basis. This includes video hosting, content recommendation, and many forms of advertising and marketing. With video, for example, the volume of data transferred represents a huge burden to non-specialist websites. (However, sites using YouTube may elect to use the ‘privacy-enhanced mode’, in which case ‘YouTube won’t store information about visitors on your website unless they play the video.’⁵) Similarly, in their current form, both content recommendation and much of online advertising and marketing require the real-time processing of user information in order to deliver targeted advertising content. The move to various forms of programmatic advertising and other highly automated forms of online marketing relies on highly complex proprietary methods and extensive data collection and analysis, and the ability of a single publisher to develop such technology on their own is minimal.

This is a hard problem, and it is clear that bringing online marketing practices in line with the GDPR will in most cases require at the very least tackling issues of user consent to the collection, sharing, and processing of personal information, not only by the large US-based technology companies that much of the public discussion around these issues have focused on, but also by the news sites and others that are intertwined with many of their products and services.

⁵ https://support.google.com/youtube/answer/171780?visit_id=0-636595692661723869-3019304114&rd=1

References

Libert, T. (2018, April). An Automated Approach to Auditing Disclosure of Third-Party Data Collection in Website Privacy Policies. In *Proceedings of WWW 2018: The 2018 Web Conference* (207–216). IW3C2 (International World Wide Web Conference Committee).

MacGregor, S and Zylberberg, H. (2018.) *Understanding the General Data Protection Regulation: A primer for global publishers*. New York: Tow Center for Digital Journalism.

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A., & Nielsen, R. K. (2017). *Reuters Institute Digital News Report 2017*. Oxford: Reuters Institute for the Study of Journalism.

Turow, J. (2011.) *The Daily You: How the New Advertising Industry Is Defining Your Identity and Your Worth*. New Haven, Conn; London: Yale University Press.

Methods Appendix

For this study, there are two main methodological considerations: developing lists of sites to study and measuring privacy impacts. These steps are detailed below.

SITE SELECTION

For each country, a list of news sites and popular sites were selected for analysis. For news, prior work conducted by the Reuters Institute for the Study of Journalism was used to identify 30 news sites in Germany, 33 in Spain, 20 in Finland, 30 in France, 31 in Italy, 29 in Poland, and 31 in the UK. For popular sites, the Alexa Web Information Service was used to determine the top 500 popular sites per country.⁶ According to Alexa, ‘traffic estimates are based on data from our global traffic panel, which is a sample of millions of internet users using one of many different browser extensions’.⁷ It is important to note that the popular sites list includes sites which users in the country visit rather than sites based in that country exclusively. For example, the Adobe website is popular in the UK but is based in the US.

⁶ <https://docs.aws.amazon.com/AlexaWebInfoService/latest/index.html>

⁷ <https://www.alexa.com/about>

MEASURING PRIVACY IMPACTS

Once the list of pages was assembled, privacy impacts were measured for the time period of 7 January 2018 to 1 April 2018. To do so, the open-source software tool webXray was used. This tool has been used extensively for academic research (e.g. Libert, 2018). As noted above, for this study webXray was configured to use the Chrome web browser. This browser was chosen as it is popular with users and it may be instrumented to run in an automated environment.

To ensure that measurement reflected what users would see in the European Union, a computer based at the University of Oxford in the United Kingdom was used. This is particularly important in the context of cookies as users in the EU have different legal protections than users in other regions, such as the US.

In addition to performing large-scale scans, webXray is also capable of performing single analyses for quickly evaluating privacy issues on a given website. On 30 April we used webXray in single scan mode to load the homepages of the 3 UK and 3 German news websites found in Figure 5. As with our larger data set, these measures were taken with the Chrome web browser and a fresh user profile with no cookies or history. These measures were likewise taken from a computer located in the United Kingdom. The figure is thus a snapshot of websites that may vary between page loads.

Acknowledgements

The authors thank Oliver Butler for guidance on GDPR issues, Reuben Binns for advice on content classification, Max Van Kleek for computational resources, and Lucas Graves for feedback on the manuscript. This research was supported by Google as part of the Digital News Initiative.

Icons courtesy of Noun Project under CC Licence

AUDIENCE ICON: Magicon (<https://thenounproject.com/search/?q=852834&i=852834>)

BROWSER NEXT ICON: Deemak Daksina (<https://thenounproject.com/search/?q=1393496&i=1393496>)

HOSTING ICON: Creative Stall (<https://thenounproject.com/search/?q=996185&i=996185>)

SOCIAL ICON: parkjison (<https://thenounproject.com/search/?q=509343&i=509343>)

SUBSCRIPTION MODEL ICON: Vectors Market (<https://thenounproject.com/search/?q=1575835&i=1575835>)

TESTING ICON: Alexa Auda Samora (<https://thenounproject.com/search/?q=1212639&i=1212639>)

ABOUT THE AUTHORS

Timothy Libert is a Research Fellow at the Reuters Institute for the Study of Journalism at the University of Oxford

Rasmus Kleis Nielsen is the Director of Research at the Reuters Institute for the Study of Journalism at the University of Oxford