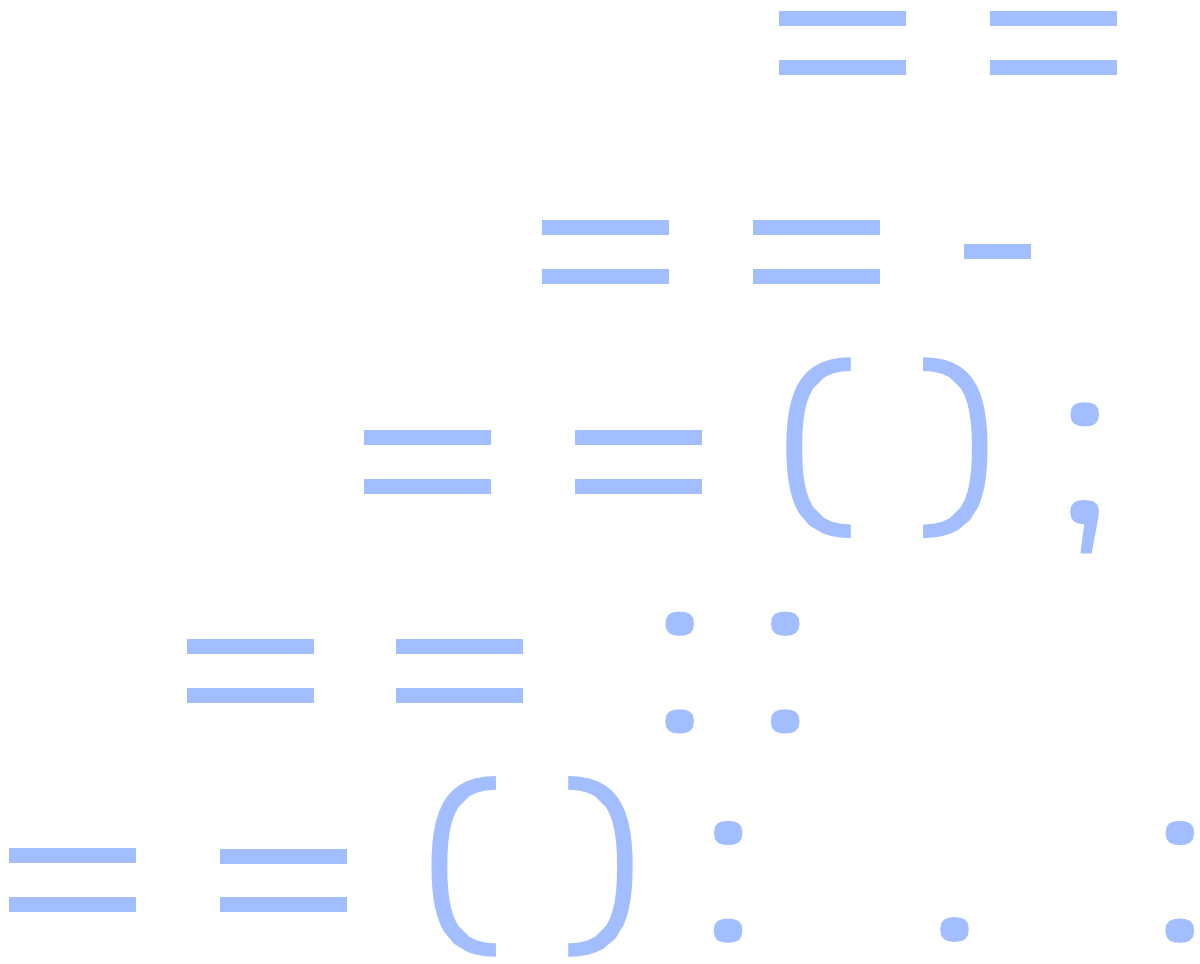
Agency for Digital Government

Report

Who is Tracking EU Citizens, and How?



December 2024

Table of Content

1. Introduction

This report aims to shed light on some of the challenges facing online privacy protection, particularly matters concerning the ePrivacy directive. It does so by conducting an analysis and evaluation of 25 of the most visited websites in the EU. The report contains three sections, each with its own subject:

1. New technologies for tracking and profiling, with a particular focus on *fingerprinting*
2. The use of Consent Management Platforms (CMPs) and the ensuing compliance
3. Who collects the data and their geographical location

The websites have been chosen due to their significance within their respective area, ranging from social media to e-commerce. The services are widely known and enjoy large user bases across European countries, making it relevant to analyze them to get a better sense of the scale and methods with which data on EU citizens are collected.

While profiling and targeted advertisement are some of the purposes for tracking users, it is important to recognize that the use of tracking technologies can be for benign or even necessary purposes. These purposes range from combating fraud or cyberattacks, to remembering the items a user puts in their cart when shopping online. While it is not the focus of this report to differentiate between purposes of the collected data, it will address its prevalence and its opaqueness to ordinary users.

2. A shift in tracking technologies

Technology is ever evolving. This also applies to technology used for tracking and profiling users. One of the technologies increasingly used for such purposes is *fingerprinting*. Fingerprinting works through the collection of various data points about the user's device such as browser type, operating system, supported fonts, list of plugins, etc. By combining these data points, a unique 'fingerprint' of a device or browser can be created and used to track a user across websites.

2.1 Results

All of the analyzed websites collect some form of data linked to browser fingerprinting. The websites made use of 203 unique fingerprinting scripts and collected data points from 33 different categories, ranging from user agent, geolocation, list of plugins, etc. These categories cover a range of various types of data[1]. It is therefore difficult to definitively conclude the sensitivity of individual data points and whether the same data point is collected numerous times.

However, reviewing the total number of unique categories that each script collects data from indicates how aggressively the user's browser is being fingerprinted. The distribution ranges from 5 for the website that collects data from the fewest categories to 28 for the one that collects from the most.

The median of the dataset is 9.5. This puts 12 of the analyzed websites above the threshold of the applied method, meaning they fall within in the medium-high group of browser fingerprinting (See Figure 1). This means that 50 pct.[2] of the websites uses browser fingerprinting in a manner that is concerning in relation to privacy, i.e. the amount of different information about the device that is collected indicates that a unique profile is being made of the user, without their knowledge or consent.[3]

[1] See Appendix A for an elaboration of the method applied
[2] Due to technical issues in the process of collecting data, Youtube.com was not analyzed in this section
[3] This is assuming that only one type of data is collected per category. This could look different, if more detailed insights into the exact type of data that is collected when a script is executed were available.

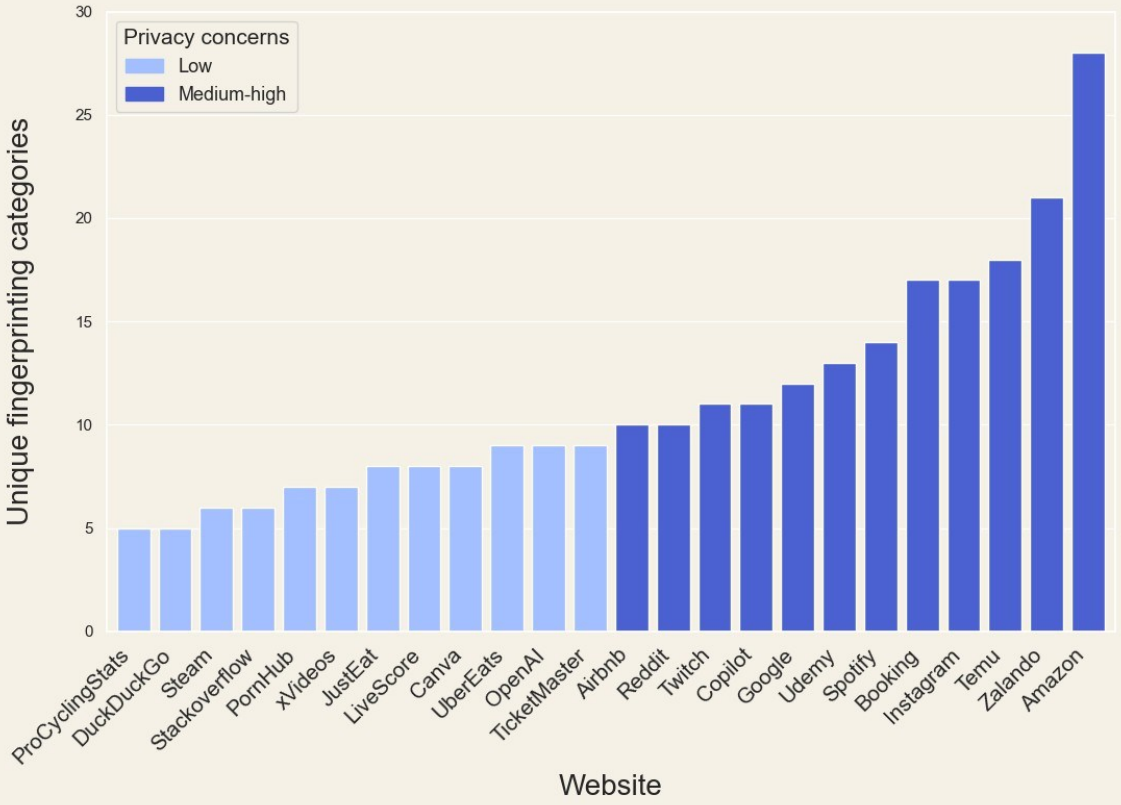


Figure 1: Unique fingerprinting categories across websites.

Although all websites employ fingerprinting technologies, reviewing the privacy policies of the websites reveals that all but two of these policies fail to mention fingerprinting, making it impossible for users to inform themselves about the tracking technologies used by the websites or to assess whether they wish to consent to the particular use of their data.

The advent of fingerprinting means that the privacy-minded user may find herself in curious dilemmas such as whether to install a useful browser extension, thus making her fingerprint more unique, against having as standardized a browser as possible in an attempt to fly under the fingerprinting radar.

For legislators, it is pertinent to question whether these dilemmas are reasonable for consumers to find themselves in, or whether more robust legislation should be put in place to protect their privacy. Furthermore, this technological shift may put pressure on the ability to apply current legal frameworks in a way that corresponds to the technological landscape, and obtain sufficient documentation for potential violations.

3. Managing consent as a business

In the face of regulatory frameworks that introduce limitations on data gathered about users without their consent, service providers have increasingly embraced consent management platforms (CMPs). This is done to collect, communicate and retain consent from users in compliance with relevant law. Today, an ecosystem with an estimated global market size of $874 million in 2023[4] has evolved to deliver consent solutions, aiming to help providers of apps and websites achieve compliance.

3.1 Results

Of the 25 websites, 10 of them manage consent via a third-party CMP provider (40 pct.), while 14 use their own consent solution (56 pct.). One of the analyzed websites did not employ any tracking registered through WebXray, nor did it have a consent solution.

9 out of the 10 websites that use a CMP provider instead of their own solution had at least one third-party service tracking the user before consent (90 pct.). By contrast, only 6 of the 14 websites with their own consent solution had at least one third-party service tracking the user before consent (43 pct.).

These third-party services are separate from the providers of the CMP, meaning the collected data are not used for the management of consent. While the exact purpose for the collection of user data might be technically necessary, we can conclude that at least some of the domains that have data sent to them before consent, are owned by adtech companies specializing in targeted advertisement.

Furthermore, when looking at which tracking technologies the websites use, only 29 pct. are cookies, while 71 pct. are other tracking technologies, e.g. fingerprinting scripts or pixels. However, an inspection of the privacy policies reveal that only 2 of the websites mention fingerprinting, while 19 mention pixels. In contrast, all mention cookies (Figure 2). This is relevant, as part of the service that CMPs fulfill, is generating cookie policies, as they are required by law, that describe the tracking technologies being used and their purpose.

[4] https://www.persistencemarketresearch.com/market-research/consent-management-market.asp

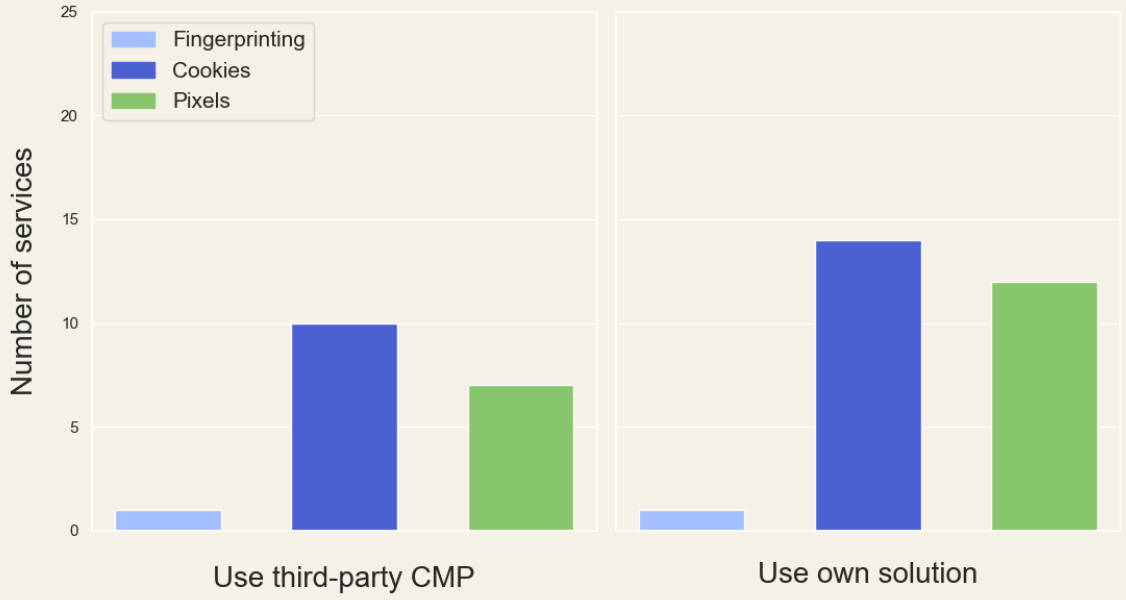


Figure 2: Number of services that mention different tracking technologies in their privacy policy.

Although we have not differentiated between the purposes of each domain call, we can nonetheless conclude that third-party domains not related to the CMP provider are prevalent even before the user has a chance to consent. While it makes sense from a cost-perspective for website owners to purchase solutions to manage consent instead of developing their own, it is not immediately clear that these solutions help website owners achieve compliance. Meanwhile, the sole responsibility for achieving compliance lies with the website owner.

4. Online privacy in a globalized world

As our lives have become increasingly digital, it has become easier to exchange information across vast distances. This allows global actors to shape the digital landscape that we traverse, e.g. by gathering data about users. While this data collection can be used to safeguard services or enable convenient options, comprehensive data collection can also be utilized to profile users and target advertisement.

The potential problems of such data collection have led a number of jurisdictions to adopt legal guardrails to ensure a sufficient level of privacy for citizens. However, as actors from a myriad of jurisdictions collect data through digital services, it has become difficult to know who has data on us, and where it flows.

4.1 Results

Figure 3 shows the number of unique trackers from their respective third-party domain owner. Although the chart only visualizes unique trackers, it is important to recognize that a tracker can be activated multiple times. Alphabet is the dominant player with a total of 55 unique trackers, twice the number of the company trailing it, Microsoft. The general trend is that there are 54 different companies tracking users across the analyzed websites.

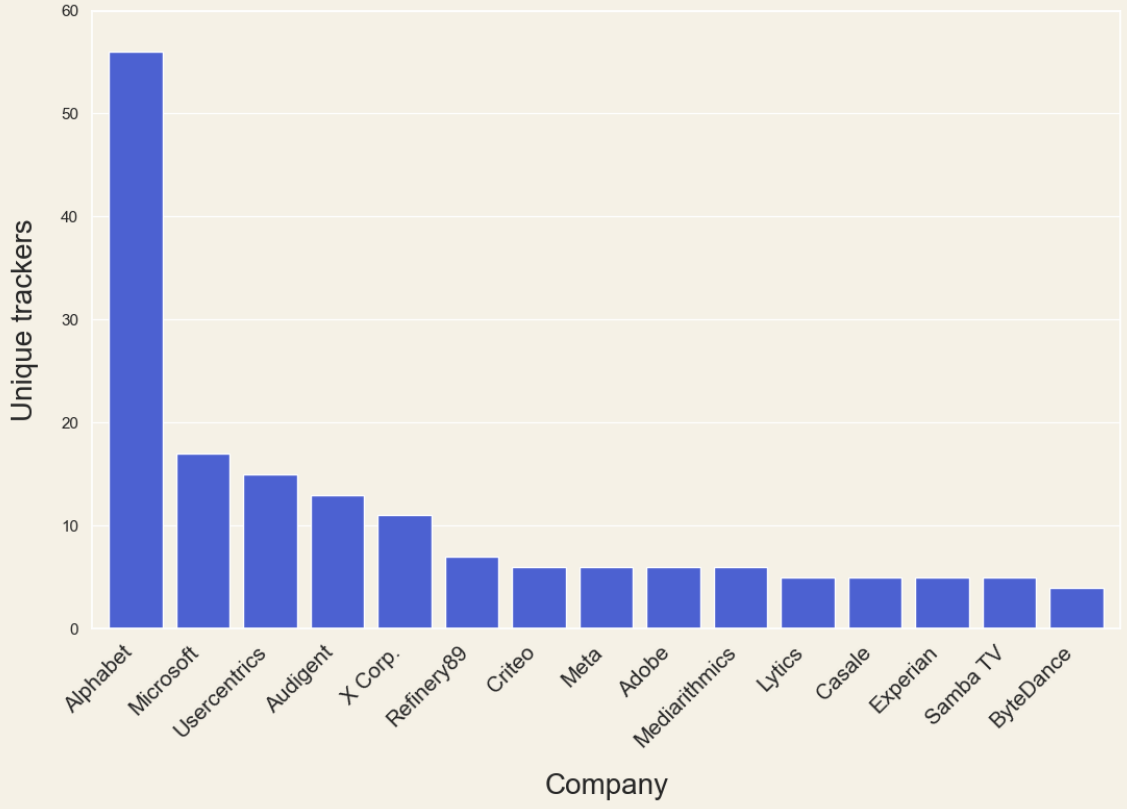


Figure 3: Number of unique trackers per parent company (top 15).

When looking at the geographical location of the services that collect data through the websites, the overarching pattern is that data is sent to 12 different destinations. Specifically, 71.8 pct. of the total amount of data collected was sent to US-based companies, followed by 8.7 pct. to Germany and 5.8 pct. to France. Thus, in comparison to the US, the remaining 11 countries account for significantly smaller percentages of the total number of trackers (Figure 4).

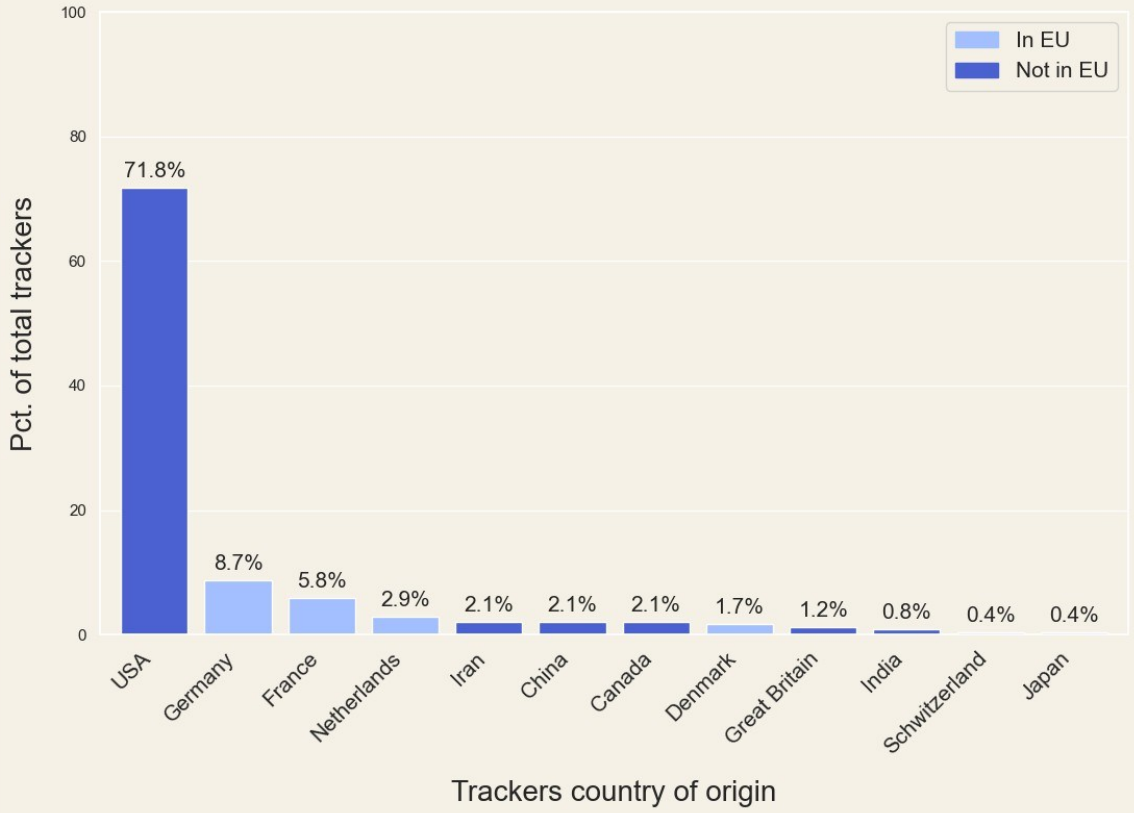


Figure 4: Percentage of total trackers divided by location of parent company.

When examining how much of the data is sent to countries outside of the EU, this accounts for 81 pct. As the US is predominantly represented, this is not surprising. However, it means that across the analyzed websites, the majority of data sent to third-parties is to companies outside the EU, often in countries with different legal frameworks governing such practices (Figure 4).

4.1.1 Result Correlations

In a previous report by The Danish Agency for Digital Government published in 2023 as *The prevalence of third-party services on Danish websites*,[5] 11.000 .dk domains were analyzed. The analysis found that 93 pct. of the websites use at least 1 service provided by a third-party based in the US. Furthermore, it revealed that the predominant third-party owner of these services was Alphabet, followed by other well-known big-tech companies such as Meta and Microsoft.

[5] https://digst.dk/media/31245/the-prevalence-of-third-party-services-on-danish-websites.pdf

Another previous report entitled *What does a free-mobile game cost?*[6], published by The Danish Agency of Digital Government in 2024 analyzed data collection by third-party services through popular free mobile games. This report found that the US, China and Israel were dominant in collecting data among the 24 analyzed apps. The report highlighted that 100 pct. of the apps sent data to the US, while 90 pct. sent data to China and Israel. Another finding was that all third-party companies collecting data from the apps were based outside the EU.

The current report supports the findings from the analysis of the 11.000 Danish websites. Looking at how many of the 25 analyzed websites employ tracking technologies from the US, this holds true for 87 pct., while 26 pct. of the websites use tracking technologies from Germany, and 13 pct. from China. Furthermore, Alphabet is the most prevalent actor across the analyzed websites, while other tech-giants also appear.

It is interesting to note the difference in prevalence of companies based outside the EU between apps and websites. While the US dominate in tracking across both apps and websites, the diversity of third-party services from countries outside the EU seems higher with apps.

These results highlight that data collection on users quickly becomes a cross-border issue, where users from one country have data sent to servers in another. While jurisdictions like the EU have legal frameworks that set guardrails to protect users' privacy, cross-border data collection by global companies make it more opaque for users where their data flows, and which steps they can take to preserve their privacy.

[6] https://digst.dk/media/31243/what-is-the-cost-of-a-free-mobile-game.pdf

5. Appendix A – Method

5.1 A Shift in Tracking Technologies

To analyze the extent of fingerprinting on the 25 webpages, the tool FPMON[7] was utilized. FPMON is an open source extension for Google Chrome. The tool registers scripts that are executed in the browser. By adding additional code to the source code, a file containing an overview of the executed scripts and a list of fingerprinting categories collected by these scripts was extracted from each website[8]. All data from FPMON was collected before any consent was given.

Some considerations about the methodic approach are necessary. Firstly, FPMON is an extension only applicable to the Chrome browser. Fingerprinting may look different in other browsers, and the results presented in this section are therefore only representative of fingerprinting on the 24 web pages when using Chrome.

Furthermore, data from FPMON does not explicitly state which data points are being collected. Rather the data points are sorted into overall categories. An example could be the category 'Storage' which covers data points such as cookies, local storage, session storage, etc. Lastly, not all potential data points used for browser fingerprinting are registered with FPMON. Therefore, there is a likelihood that some of these are not included in the final data used for the analysis.

To determine whether or not a particular fingerprinting activity is extensive, thus indicating some form of profiling or tracking taking place, we have looked to the creators of FPMON's own categorization when using the tool. In their own use of FPMON, a threshold is determined by calculating the median of the total number of fingerprinting categories each website collects data from. The total number of categories below the median are categorized as 'low' while those equal to or above are divided into 'medium' or 'high', Whether a website falls into the former or the latter depends on the number of sensitive features collected from each category. As we do not have access to the specific types of data collected, websites are categorized as 'medium-high' if they are equal to or above the median.

5.2 Managing Consent as a Business & Online Privacy in a Globalized World

The primary data collection tool for this section is WebXray[9], which is designed to analyze website-tracking practices. WebXray opens each website for approximately 15 seconds to identify tracking technologies that are activated before any consent is given. By manually categorizing the extracted data, and analyzing the websites using browser-enabled developer tools, we can determine whether a website has a CMP solution as well as whether a third party supplies it.

It is essential to consider the limitations of WebXray. Websites often show dynamic content depending on geographic location or user-specific factors. To keep things consistent, this analysis focuses solely on .com domains, avoiding national variations. However, due to geo tracking and the fact that we accessed the websites from Denmark without using a VPN, results may vary if visiting them from

[7] https://fpmon.github.io/fingerprinting-monitor/
[8] Due to unforeseen circumstances data was not collected from Youtube
[9] https://github.com/thezedwards/webXray

elsewhere. Moreover, the 15-second window might not capture the entire bundle of tracking technologies due to delays in their activation. However, because cookie banners are typically the first thing users should be exposed to when visiting a website, WebXray is well suited for the scope of this analysis. Lastly, WebXray visits websites in a so-called headless browser, which can be identified by the websites, who may opt not to execute certain tracking when a visit is registered to employ a headless browser.

By combining data from WebXray with data on whether the 25 websites manage consent via solutions built in-house or with the aid of a third-party provider, we can begin to identify trends and differences between these approaches.

Using the WebXray data from previous sections of the report, we can identify the owners of third-party domains that are called as a user visits one of the 25 websites. This enables us to map the ownership structure of the companies who own the domains, in addition to their location. In cases where the domain points to a subsidiary, its parent company have been ascribed ownership of the domain. For example, we have determined Alphabet as the owner of doubleclick.com domains. To determine this, we relied on data from Crunchbase[10] and desk-research using common search engines.

Additionally, the privacy policies of the websites have been manually reviewed in order to determine whether or not specific tracking technologies are mentioned.

[10] https://www.crunchbase.com/

6. Appendix B – Analyzed websites

The 25 websites analyzed for this report are the following:

Social media
1. Reddit
2. Instagram

Entertainment and gaming
3. Spotify
4. Youtube
5. Twitch
6. Steam

Sports
7. Livescore
8. Procyclingstats

Productivity and digital tools
9. ChatGPT
10. Copilot
11. Canva

E-commerce
12. Zalando
13. Ticketmaster
14. Temu
15. Amazon

Travelling
16. Booking
17. Airbnb

Pornographic
18. Pornhub
19. Xvideos

Search engines
20. Google
21. DuckDuckGo

Education
22. Udemy
23. Stackoverflow

Delivery platforms
24. Just-eat
25. Uber Eats

Agency for Digital Government