Department of Computer Science

Submitted in part fulfilment for the degree of
MSc in Cybersecurity.

"Accept All"
The Cookie Banner Landscape in Greece & the UK

Georgios Kampanos

16 September 2020

Supervisor: Dr Siamak Shahandashti

Acknowledgements

Thank you to my studies and project supervisor, Dr Sia Shahandashti, for providing guidance and feedback throughout the year and especially during this project.

Thank you also to my girlfriend Anthi, for putting up with my complaining and supporting me throughout my studies. You have been amazing and without you, this wouldn't have happened.

Abstract

The increase in internet usage and online services such as internet banking and e-commerce has brought with it extended online-tracking and data gathering. The European Union and the UK have reacted to this new threat by passing laws such as the GDPR and the Data Privacy Act 2018. Such legislation forces websites to disclose their tracking activities and the parties that have access to the collected information.

In response to the legislation, websites implemented Cookie Banners – pop-ups that appear when a user visits a website for the first time. Since Cookie Banners are commonplace, this project has aimed to broaden the understanding of the Cookie Banner landscape, how they affect users and what type of privacy options they offer.

Specifically, this project looked at cookie banners in Greece and the UK since they are both governed by Data Protection laws but they also differ in language and size. More than 14,500 websites were surveyed across both countries. This was done by using OpenWPM [1], a popular web-privacy measurement framework, which was extended to detect and store Cookie Banners. Furthermore, for the purposes of this study, a number of novel methods and techniques were developed that were used to identify popular websites and sanitise and normalise the collected data.

This project collected more than 7,500 cookie banners and over 15 million datapoints from OpenWPM. The results show that although 52% of websites have implemented Cookie Banners, more than 12% of the sample does not have one and therefore, not complying with the law. Furthermore, it is evident that websites use a number of dark patterns to nudge users towards privacy-intrusive choices. Specifically, only 5% of websites across both countries have direct opt-out links. Moreover, the majority of websites nudge users into accepting cookies since they "improve their browsing experience".

Statement of Ethics

- **Aim**: This project aims to determine the number of privacy options that a visitor is given when visiting Greek & UK websites.

- **Methodology**: OpenWPM is an open-source web privacy measurement framework that allows users to collect privacy information from websites and is already been used in several other studies. Using OpenWPM, a large number of websites can be crawled and detect and store their cookie notice in a database and then perform analysis on that stored notice (e.g. how many websites offer an "opt-out from third-party tracking" option).

- **Participants**: There are no participants in this study.

- **Data collection**: Only the HTML from the cookie/privacy notices is collected. This information cannot be used to deteriorate the security of that website (e.g no functional code is collected to identify bugs). However, the robots.txt file is respected and websites that "Deny" crawlers or have no robots.txt at all are not scraped. Furthermore, this project makes best efforts to respect the "Terms & Conditions" of the websites in our list by ensuring that they do not contain phrases such as "for personal use only". If such exclusionary terms are found, then the website is not crawled. Specific details of the robots.txt and Terms of Service compliance are discussed in detail in Chapters 3 and 4.

- **Copyrights**: OpenWPM simulates "real" users and therefore, visits websites using a consumer browser (Firefox). Therefore, no additional information is downloaded and no data is maintained for longer than required.

- **DoS**: OpenWPM visits the websites from the given list only once, gathers the required data and then drops the connection of the website. Furthermore, all additional data analysis will be conducted offline.

Therefore, no websites are going to be DoS during the survey.

Contents

List of Figures

List of Tables

1 Introduction

EU citizens engage in a wide range of activities online which include getting news, banking, social media, shopping and many more. According to the Digital Economy and Society Index (DESI), in 2019 Sweden, Denmark and the UK have the most advanced digital economies. On the other hand, Romania, Poland and Greece score the lowest in DESI [5].

Even though 85% of Europeans used the internet in 2019, the COVID-19 pandemic is expected to contribute to the rise of the number of internet users as well as interactions and services [6]. With the increase in internet usage, there has been a rise in the amount of information collected by websites, such as location data [7], as well as online tracking with the use of Third-Party Cookies (TPs) [8].

Thus, the rapid growth of the internet and its online services has brought a new set of challenges and risks to citizens and governments alike. While the EU and the UK have introduced legislation in an attempt to mitigate and control those risks, the privacy of internet users at risk on a daily basis.

1.1 Legislation

The EU and the UK have implemented laws in an attempt to protect their citizens from mass-scale data collection. Furthermore, such legislation gives the ability to governments to bring companies to justice for mishandling personal information, as seen from the Cambridge Analytica scandal [9].

European Union

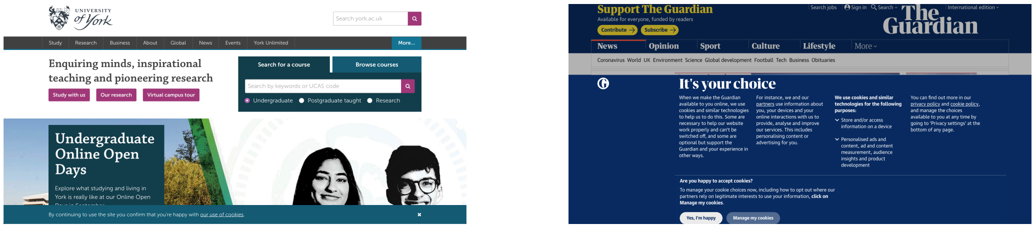
The General Data Protection Regulation (GDPR) is a data protection and privacy law in the European Union that came into effect in May 2018. The primary goal of the GDPR is to give the ability to citizens to control their personal information online. For instance, it allows users to request their data held by companies and also their data erased. Furthermore, websites are required to disclose if they have implemented data collection mechanisms, as well as how long the data is going to be retained for and whether they are going to be shared with third parties or outside the EU [10].

United Kingdom

The Data Protection Act of 2018, is a UK parliament Act that aims to update the data privacy laws in the United Kingdom [11] (Data Protection Act 2018). Furthermore, It has been written to complement the GDPR but it is not limited by the GDPR's provisions.

1.2 Cookie Banners

Cookie banners are pop-ups that appear when users visit a website. They can either take the full screen, forcing users to interact with them before continuing to the content or appear as small bars allowing users to bypass them. Figure 1.1a, depicts the "full-screen" cookie banner of The Guardian (`https://theguardian.com`) and Figure 1.1b shows the cookie notice implementation of the University of York's (`https://york.ac.uk`) which does not force the user to interact with it.



Enquiring minds, inspirational teaching and pioneering research

Explore what studying and living in York is really like at our Online Open

By continuing to use the site you confirm that you're happy with our use of cookies.

It's your choice

When we make the Guardian available to you online, we use cookies and similar technologies to help us to do this. Some are necessary to help our website work properly and can't be switched off, and some are optional but support the Guardian and your experience in other ways.

For instance, we and our partners use information about you, your devices and your online interactions with us to provide, analyse and improve our services. This includes personalising content or advertising for you.

We use cookies and similar technologies for the following purposes:

- Store and/or access information on a device
- Personalised ads and content, ad and content measurement, audience insights and product development

You can find out more in our privacy policy and cookie policy, and manage the choices available to you at any time by going to 'Privacy settings' at the bottom of any page.

Are you happy to accept cookies?

To manage your cookie choices now, including how to opt out where our partners rely on legitimate interests to use your information, **click on Manage my cookies**.

(a) The "full screen" cookie banner shown to The Guardian's visitors.

(b) The "bypass-able" cookie banner shown to York Uni's visitors.

Figure 1.1: Two different cookie banner implementation approaches.

As the GDPR states, when websites are storing tracking cookies on a user's browser they are required to inform the visitor on how their data is going to be used and who is going to have access to them. Thus, websites implement cookie notices in order to inform users on their cookie policies and allow visitors to control their privacy settings via "privacy options" offered on the cookie banners e.g: "Accept all" or "Decline".

Cookie notices can be a valuable tool for EU and UK and citizens to take control of their privacy online. However, they can also provide a powerful platform for websites to steer users towards privacy-intrusive choices in order to collect even more information on their visitors.

1.3 Motivation

It is evident that the ubiquitous nature of cookie banners means that they are rapidly becoming part of a user's daily web browsing experience. Especially after the GDPR came into force in 2018, more and more websites have added such notice.

Since users increasingly have to deal with those privacy notices this project aims to provide a better understanding of cookie banners and the privacy options that they provide. More specifically, the research will focus on the types of cookie banners that users have to interact on a daily basis, whether visitors have a choice on whether they are being tracked and how the wording on those cookies notices may affect users' perception of privacy and tracking.

1.4 Aim

This project focuses on the cookie banners in Greek as well as UK websites. These countries were chosen for 2 reasons. Firstly, they both adhere to very similar data protection laws, namely the GDPR and the Data Protection Act of 2018 and therefore, it is expected that most websites in these countries will have cookie notices. Secondly, both countries vastly differ in language and population size. Thus, potential differences in how the two populations experience the internet and the cookie banners are going to be highlighted.

The following list summarises the main goals of this project:

1. **Users**: Develop a comprehensive understanding of the cookie banner landscape is useful not only for users but also for developing technologies that help users manage their privacy rights;

2. **Governments**: Assist policymakers to have a better idea about the level of compliance to privacy laws by websites as well as the dark patterns that users encounter on a daily basis.

1.5 Research Questions

In order to understand cookie banners better and how they affect the users' internet experience, this project sets 7 research questions. These are summarised in the following list:

RQ1: What is the prevalence of cookie banners in popular websites across Greece and the UK?

RQ2: How many privacy options do cookie banners provide on average?

RQ3: How many cookie banners offer their users a direct "opt-out from tracking" option?

RQ4: How many cookie banners do not offer any option at all and inform their users that by "using this website, they agree to Third-Party Cookies and tracking"?

RQ5: What is the most common privacy option provided by the cookie banners?

RQ6: How many cookie banners allow their users to manage their privacy settings and control which vendors track them?

RQ7: What is the average length of the cookie banner privacy text and what are the most common terms that are used to inform users about the use of cookie banners?

1.6 Contributions

In order to answer the research questions set above, a number of novel methods and software was developed and a plethora of data was collected throughout the course of this project. All the contributions made are summarised in the following list:

1. Developed an automated method of scraping and collecting cookie notices on a large scale, using OpenWPM;
2. Conducted a more comprehensive study in Greece and the UK and collected a significantly larger dataset compared to other similar studies such as Habib et al.;

3. Make the tools and data available so that similar research can be undertaken by other researchers in different countries.

2 Related Work

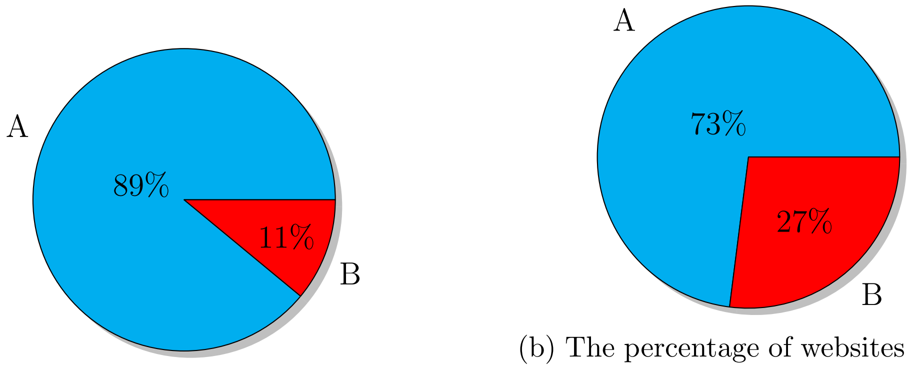
This chapter introduces previous work that has been undertaken on the subject of cookie notices and consequently, has influenced this project as well. Within the body of past work, 5 distinct subcategories have been identified and discussed further. These subcategories are summarised in the following list:

1. Cookie banners and the privacy options that they offer to their visitors;
2. Whether user location impacts cookie banners and privacy policies;
3. The user interface (UI) of those cookie banners as well as dark patterns that websites employ to influence a user's privacy choices;
4. The effect that the GDPR had on cookie notices and how the privacy landscape changed;
5. The prevalence of online tracking using cookies regardless of whether the user has agreed to be tracked or not.

2.1 Cookie Notices & Privacy Options

Nowadays, cookie banners and privacy notices can be found in the vast majority of websites. Fundamentally, they are there to notify users that the website is using cookies, for advertising or other purposes, but also give users a means to manage their privacy settings or seek further information.

However, not all websites provide their users with direct links to opt-out from tracking and instead are hiding those functions away from the cookie notices. This was researched by Habib et al. [12] who conducted a 150-website analysis that aimed to determine the privacy mechanism and options available to users of those websites. Firstly, they wanted to know what are the options available to users in regards to email communication and targeted advertising. Secondly, they aimed to determine the ways that those websites present their privacy options to the users. Using the Alexa worldwide website rankings (`https://www.alexa.com/`), the authors used two thresholds for dividing and categorising those websites. Those were the top websites (ranks 1 - 200), middle websites (ranks 201 - 5,000) and finally bottom websites (ranks > 5,000). Interestingly, the researchers found that privacy options are frequent within the surveyed websites. Specifically, they found that 89% of websites offer email communication and targeted advertising opt-out options and 74% of websites "had at least one data deletion mechanism", which was higher than what other similar studies had found, as shown in Figure 2.1a. Yet, direct opt-out links are found in only 27% websites as depicted in Figure 2.1b and therefore, multiple steps are needed for a user to be able to find a use that link. Moreover, they found that privacy policies contained missing, misleading and even unhelpful information. For instance, in 6 websites the text referred to opt-out but "that opt-out did not exist". Moreover, the authors noted that in 15 websites, there were broken links or opt-out mechanisms and therefore, users could not opt-out even if they chose to.



(a) The percentage of websites offering opt-out options (A) and the websites that do not offer one at all (B).

(b) The percentage of websites offering direct opt-out options (B) and the ones that require multiple steps (A).

Figure 2.1: The opt-out options offered by the websites in Habib et al. sample.

Regardless of whether websites offer cookie notices or not, websites in Europe have to ensure that their cookie notices follow the rules set out by the GDPR. Nonetheless, it is true that a number of websites have yet to adapt to the new rules, risking fines but also their users' privacy. This is supported by Nouwens et al.'s work [13] who set out to find whether Content Management Platforms (CMPs) and their consent pop-ups adhere to the rules set out by the European Union. More specifically, the researchers aimed to determine how prevalent non-compliant design elements on cookie notices are and how those user interfaces affect the users' choice in regards to their privacy. They surveyed 10,000 UK-based websites but only 680 contained a CMP that they were able to successfully scrape. The researchers found that 32% of their sample had "implicit consent" privacy notices. This means that users were agreeing to be tracked because they were using the website which the GDPR does not allow. Interestingly, they found that CMPs make rejecting trackers more difficult than accepting it. Only "12.6% of sites had a *reject all* button accessible with the same or fewer number of clicks as an *accept all* button". Moreover, the *accept all* button was never hidden in contrast to the *reject all* button that required additional steps to be found. Furthermore, 56.2% of the websites had pre-selected optional vendors and purpose/categories. Their findings are depicted in Figure 2.2.

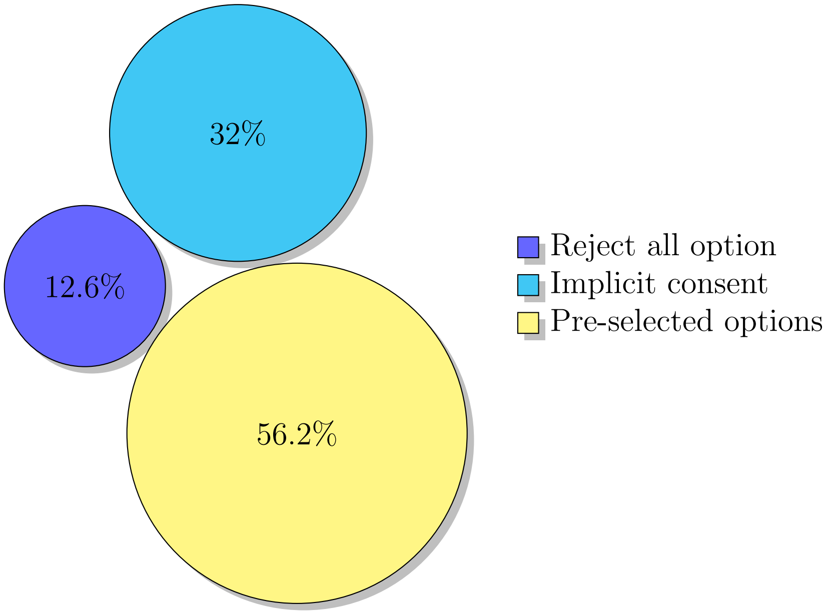
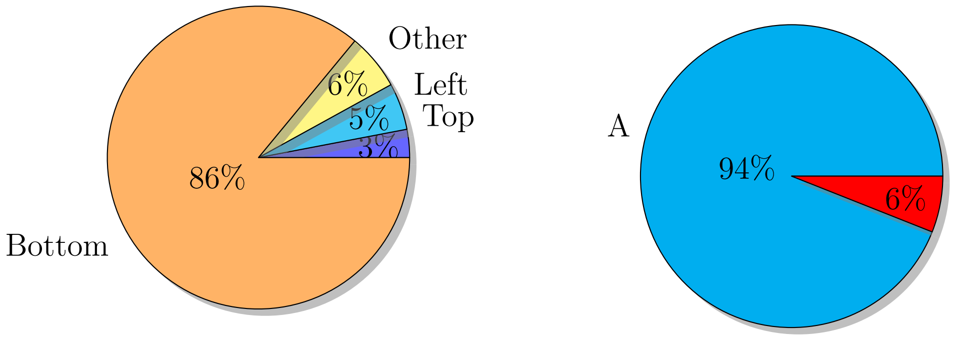


Figure 2.2: The options offered by CMPs as shown by Nouwens et al.

In addition to cookie banners, websites also have specific pages that list their privacy policies. There, websites explain how data collected from users is going to be used and who it is going to be shared with.

While this seems positive in allowing users to be more informed about their privacy, these pages are often hidden or hard to read by an average user. This was found by Jensen and Potts [14] who conducted research in order to determine the different aspects of privacy policies on US websites. More specifically, they aimed to determine how readable is the content of those privacy policies as well as how accessible they are. Their data was divided into two subcategories that included high-traffic websites and low-traffic ones. The first set was used in order to simulate how most internet users view those policies on a daily basis and the second set was used to examine the effect of regulatory efforts to improve privacy policies. With regards to accessibility, the researchers found that 86% of the websites have a link on the bottom, 3% on the top and 5% on the left of the page that takes users to a privacy policy page. In total, 94% of the websites offer a direct link to the privacy policy page while the remaining 6% require users to go through an intermediate page to access the privacy policy. Interestingly, 8% of the websites obscured the privacy policy link through some sort of formatting and 27% of the websites offered that link in a smaller font, compared to the rest of the text on the website. Regarding the readability of privacy policies, the researchers determined that only 6% of the websites were "accessible to 28.3% of the Internet population with less than or equal to high school education". Furthermore, 13% of the policies required the equivalent of post-graduate education for a visitor to comprehend. Finally, content-wise, they found that 13% of the websites did not explain to the user how they will communicate changes to their privacy policy to the visitors. On the other hand, 19% of the websites offered to notify the users by email if changes were to occur to their privacy policy while 69% "required users to check the policy page periodically".



(a) The percentage of the privacy policy link position offered by a website.

(b) Percentage of websites offering direct (A) and indirect (B) links to their privacy policy page.

Figure 2.3: ?

Even though cookie banners and privacy policies are a reality, the question still remains as to what users think about them, since they are the ones impacted by them the most. Borgesius et al. [15] created a survey in order to try and understand what users think about those cookie walls and whether they think that they are fair. Furthermore, the researchers aimed to determine if people thought it was fair to exchange their personal information for free content (e.g. news articles). The survey was conducted in the Netherlands and it was taken by the total number of 1,235 people. The researchers found that 60% of the participants thought that tracking walls were not fair or acceptable. This included all types of websites including shopping, news and health. Moreover, 64% of the participants felt that "trading personal data against the use of a *free*" website is unacceptable. However, only 44.9% of the participants think that is not acceptable for a free website to contain ads.

2.2 Cookie Notices & Location

Privacy and tracking laws can vary from country to country. For instance, European privacy laws, such as the GDPR, do not apply in the USA and vice versa. However, when users from the USA visit a Spanish website, that website does not have to follow rules set out by the USA. Their cookie banner and privacy policy can follow the rules of their country of origin. This was found by research done by Eijk et al. [16] who looked at whether user location impacts cookie notices or if websites follow the rules of their country of origin. They hypothesised that in order for websites to simplify their decision process, they follow the rules of their main target audience, for instance, their Top Level Domain (TLD). Using web scraping and a list of crowd-sourced CSS cookie banner selectors [17], they set out to identify and measure the cookie banners found on websites. In total, they surveyed 1,543 European, American and Canadian websites. They were able to detect a cookie notice on 40% of the websites that they surveyed with "a median of 4 Third-Party Cookies" being stored by each website. However, the researchers could not find any evidence to suggest that the cookie notices were being affected by the user's location. Thus, their initial assumption that the websites follow the rules based on their country of origin or TLD holds.

Similar research was undertaken by Fruchter et al. [18] However, they investigated whether the number, as well as the type of trackers, are different between countries. Furthermore, they set out to determine whether the trackers are impacted because of the regulatory model that exists in each country. Four distinct types of regulatory models have been identified which are summarised in the following list:

- "**Comprehensive**": Includes countries that view privacy as a fundamental human right. They require organisations and companies to protect users' personal information by limiting how much data is being collected and used. The European Union has adopted this model;
- "**Sectoral**": Governments enact laws that may target a specific in-

dustry e.g. the financial sector but they do not provide fundamental protections on privacy. The sectoral model is adopted by the United States;

- "**Co-regulatory**": Organisations and industries have to self-regulate and develop their privacy-policies for data protection and privacy. This model has been adopted by Australia.;
- "**Mixed / no-policy**": This model is used by countries that do not protect privacy, such as China, or use a combination of the previously mentioned models.

The researchers, using web crawling, aimed to determine the amount of web tracking within different countries that represent 3 different models (comprehensive, sectoral and co-regulatory). Those countries are Germany that has implemented the comprehensive model, the United States and Japan which both represent the sectoral model and finally, Australia that uses the co-regulatory model. Interestingly, the researchers found that when visiting websites from the US, those websites stored more Third-Party Cookies and made more HTTP requests than when visiting from somewhere else. However, they noted that "a website's country of origin or a server's physical location has even more impact than a user's geographic location". Therefore, this is an indication that the website's country of origin, instead of the users', matters when it comes to the level of tracking. Furthermore, the researchers showed that there is approximately 2% more HTTP request made by trackers compared to advertisements. They determined that this is particularly problematic since the trackers do not have a visual element and therefore, a user might have a false sense of privacy when browsing those websites. They also made comparisons between two countries within the same model, namely the United States and Japan that have employed the sectoral model. They found that the United States "showed significantly greater" levels of tracking-related requests and cookies compared to Japan. They attributed this fact to potential cultural differences or different types of popular websites in each country that may indicate a different business model (e.g. not interested in making money from advertisements). In conclusion, the researchers were not able to draw any conclusions about the regulatory models themselves. That is because they were unable to find "interesting results when examining each individual country". Thus they were not able to find a variation or a pattern that could be explained beyond what can be explained by the model employed by the country.

While a user's location might not change the privacy options offered by cookie banners, websites are required to follow the rules set by their country of origin. After the GDPR came into force, websites had to adapt to a new set of rules set by the European Union.

Matte, Bielova and Santos [19] investigated whether European websites and their cookie notices adhere to those new rules set by the GDPR and ePrivacy directive. They focused on whether banners actually respect the user's choice or whether they register a positive consent regardless of the visitor's choice. Furthermore, they looked if banners nudged (i.e. have pre-select privacy options) the user to accept everything. In total, they surveyed 28,257 websites but only 1,427 of these had a cookie banner that the researchers were able to experiment with. They were able to detect 4 different types of violations. Firstly, they found that 141 websites registered an affirmative consent without before the user had performed any actions and 38 websites offered no "opt-out" option at all. Similarly to the finding by Nouwens et al. discussed earlier, they observed that at least 50% of the websites on their set have pre-selected privacy options on their privacy notices "nudging" users towards privacy-intrusive choices. Finally, the researchers found that at least 27 websites did not respect the user's choice even though they declined to be tracked by cookies. Table 2.1, summarises the violations that were committed by websites as shown by Jensen and Potts.

Violation	**% of websites**
Registering affirmative consent	9.8
No opt-out option	2.6
Pre-selected privacy options	50
User choice not respected	1.8

Table 2.1: The violations committed by websites and how prevalent they are within Jensen's and Potts' dataset.

2.3 User Interface of Cookie Notices

As discussed in the sections above, cookie notices are a reality for every internet user whether they think they are fair or not. However, websites can exploit certain aspects of the User Interface (UI) of those cookie notices in order to drive users to make privacy-intrusive choices. For instance, since most users think that cookie banners are a nuisance, websites can have pre-selected privacy-intrusive options betting that users will click on the "accept all" button just to get rid of the notice.

Those dark patterns and practices were investigated in detail by Utz et al. [20] who conducted research in order to study the design properties of cookie banners in a number of different websites in the European Union. More specifically, they performed three different experiments. Firstly, they looked at the positioning of the cookie notices and whether the location on the website affects the visitor's consent decisions. Secondly, they focused on whether the number of options that the cookie notices provide affect users and to what extent. Thirdly, they looked if the existence of a privacy policy link or the choice of words (e.g. Technical vs Non-Technical wording) in the cookie banner affects the visitor's consent decision. Initially, they developed a list of the top 500 websites for each European country, which "yielded a list of more than 6,000 unique domains". Then, using a European IP address, they took screenshots of the homepage of each website on the list mentioned before. Finally, they manually inspected the screenshots to verify that they contained a cookie banner. The researchers looked for a number of user interface attributes such as size, position, whether it is blocking the user from using the website before choosing an option as well as the type of choices offered to the visitor (e.g. No option, binary, confirmation only etc.). The first looked at the effects of the banner position on the users. It showed that when cookie notices were placed on the bottom-left of the website, they "received the most interactions". The researchers noted that 33.1% of the users interacted with those notices regardless of their device or choice made. The second experiment also showed that nudging a user and pre-selecting their choice had a significant impact on the visitor's final choice. The final experiment that was concerned with the privacy policy text showed that using technical language (i.e. "cookies" instead of "your data") had very little impact on a user's final choice. This was probably due to the fact that most users do not pay attention to the notice and just accept the default or pre-select option that the banner gives them. These findings indicate that the position of the banner and the given options have a bigger impact on a user's choice compared to the notice language or to the privacy information.

A similar survey was undertaken by Kulyk et al. [21] who aimed to determine the effect that cookie banners had on users when used as privacy notices. The researchers focused on 3 main questions. Firstly, they wanted to establish what users think of cookie notices when seeing them on websites. Secondly, they wanted to know the users' reactions when viewing and interacting with privacy notices. Thirdly, they wanted to determine the factors which influence user decisions in regards to their surfing behaviour, when they see a cookie disclaimer. In total, 150 people participated in the study, including 73 females, 75 males and two participants who did not specify their gender. The survey results have been classified into 5 distinct categories which are summarised in the following list:

1. **Disturbance**: The researchers found that a large number of participants felt annoyed by the cookie notices and considered them to be a disturbance when browsing the web. One participant said: "As these messages appear constantly, I find them to be disruptive and annoying";
2. **Privacy concerns**: A number of participants felt concerned about their privacy when they saw the cookie banners with one participant saying that "I feel observed";
3. **Habituation**: Interestingly, the researchers noted that because of the prominence of the cookie disclaimers, a large number of participants felt they were used of them and therefore, did not pay attention to them;
4. **Lack of information**: The researchers found that participants felt that they were not informed enough to understand the consequences of cookies in regards to their privacy. They believed that there was a need for more detailed banners that included information such as how their information is collected and is used by the websites. One participant specifically said: "It is unpleasant to me, as I do not know exactly what it means to allow cookies, and what consequences it has for me";
5. **Misconceptions**: The researchers found that a number of participants had misconceptions regarding what cookies are and what the consequences of cookie use are. For instance, one participant said: "Maybe I have a feeling that I am attacked by a virus".

Unfortunately, the practices described in the above list are also employed by big tech companies such as Google and Facebook. Since these companies rely on selling personalised advertisements in order to generate a profit, they try to discourage users from opting out of tracking. This was investigated in detail by the Norwegian Consumer Council [22]. They looked at whether user interfaces of cookie notices and privacy settings, provided by Google, Facebook and Microsoft's Windows 10, discourage users from making privacy-aware choices. They found that all 3 companies provide default settings that are considered privacy intrusive and the cookie notices contain misleading wording. On the contrary, "privacy-friendly" options that allow users to opt-out from tracking, require multiple steps to find. The

Council noted that “design, symbols and wording that nudge users away from the privacy-friendly choices” are prevalent in cookie banners from all 3 companies. Interestingly, key information is sometimes omitted and instead the wording in these notices is written in a way to compel users into making privacy-intrusive choices. Furthermore, the researchers found that both Google and Facebook “threaten users with loss of functionality or deletion of the user account” unless the user agrees to those privacy-intrusive settings. The Norwegian Consumer Council’s findings on the dark pattern employed by the big tech companies are summarised in Table 2.2.

	Facebook	**Google**	**Microsoft**
Privacy intrusive defaults	✓	✓	✓
Hard to find opt-out buttons	✓	✓	✓
UI promoting privacy intrusive choices	✓	✓	✓
Loss of functionality warnings (after account deletion)	✓	✓	

Table 2.2: Dark patterns that big tech companies employ, as observed by the Norwegian Consumer Council.

Since tracking is rife on the internet, a number of services and tools have been created to help users protect their privacy. These tools include web browser built-in plugins as well as third-party tools such as AdBlock Plus [23], a popular plugin that is used for ad blocking on websites.

While these tools are doing an excellent job at stopping tracking, they can be hard to set-up and use by an average user and therefore, they can be rendered useless. On this topic, Leon et al. [24] researched the usability of privacy-focused tools that limit Online Behavioural Advertising (OBA). In a lab setting, they interviewed a number of participants and recorded behavioural patterns when installing and using privacy tools such as browser plugins or in-browser settings. The researchers surveyed 9 different tools from 3 different categories. Their list included 3 opt-out tools, two built-in browser settings and four blocking tools. Significant flaws were observed in all 9 tools tested in the laboratory setting which made it hard for regular users to protect their privacy while browsing the web, even if they wanted to. They found that the vast majority of users lack “sufficient knowledge” in regards to privacy and tracking technology and therefore, do not know how to use privacy tools properly and often choose not to use them. Furthermore, the researchers noted that it is hard for users to keep up as trackers, privacy tools and technology are constantly changing. Therefore, it is challenging to provide “easy-to-use tools that give users meaningful control without interfering with their use of the web" but for users, it is hard to make privacy-aware choices "without breaking desired website features".

2.4 Cookie Notices Before and After the GDPR

The GDPR came into force on 25 May 2018 and set out rules on how websites should treat people's privacy, allow them to manage and delete their data as well as inform users that they are being tracked with the use of cookies. Therefore websites, and their cookie banners, had to adapt to the new rules in order to avoid penalties and fees.

Unfortunately, a significant number of European websites never made the transition to the post-GDPR era and are still making it difficult for users to opt-out of tracking or delete their data. This was extensively investigated by Sanchez-Rola et al. [25] who looked at whether websites respect a user's choice to opt-out from tracking after the GDPR went into force. More specifically, they visited 2,000 popular websites from all around the world and tried to refuse tracking, when the option was available, while a custom plugin was collecting the number and type of cookies stored in the user's web browser. The research aimed to determine whether users could control cookie tracking after the GDPR went into force. Although their sample was quite small and used a manual method of analysing their data, the researchers found that approximately 92% of the websites that they visited, they track users even if the user does not give their consent. The authors noted that this happens "even before showing any banner about cookie policies". Furthermore, the researchers found that only a few websites allow users to opt-out from tracking. More specifically, only 4% of the surveyed websites had a clear "reject" button and in some cases, the users had no options at all (i.e. the website informed them "by using this website you agree to the use of cookies"). Similarly, only a few websites have an "Options" interface that allows users to control the type of cookies that are going to be stored on their web browsers. The researchers went further and showed that when users reject tracking, it is often ineffective and the websites continue tracking them. They noted that "in most cases, the number of cookies set by the server remains the same or even increases". Only 2.5% of websites erase the whole set of cookies that they set after the visitor has opted-out. Figure 2.4 summarises the findings by Sanchez-Rola et al.

Similar research was conducted by Degeling et al. [26] who looked at changes that occurred at privacy policies as well as cookie banners on European websites before and after the GDPR came into force. In contrast

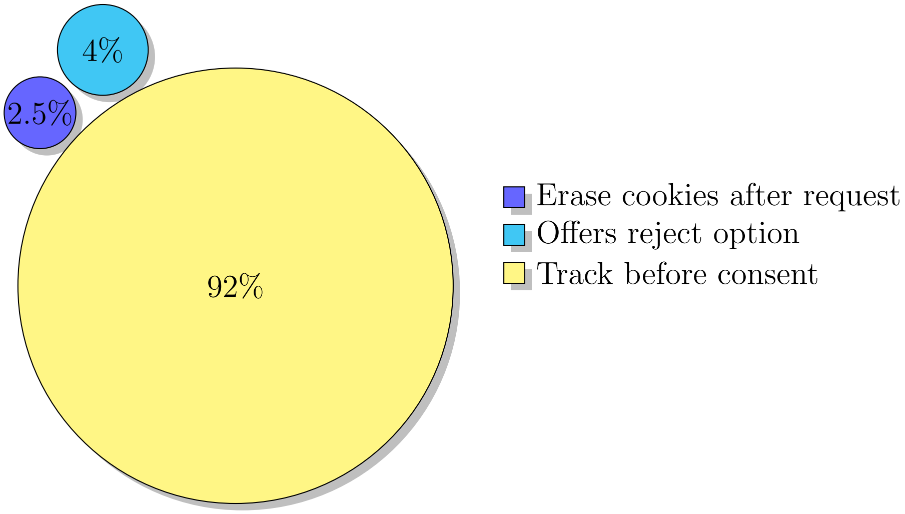
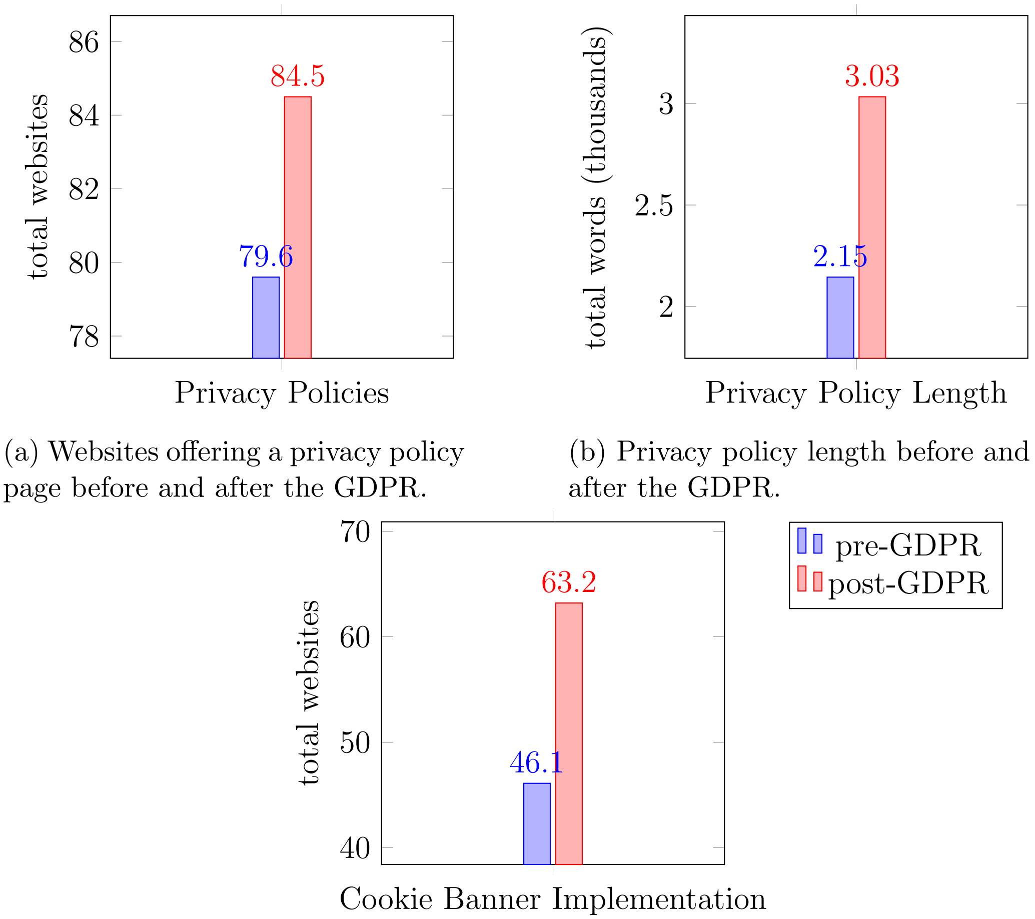


Figure 2.4: Tracking behaviour of websites as observed by Sanchez-Rola et al.

with Sanchez-Rola et al. work, Degeling et al. had an automated method of gathering data which also allowed them to analyse a larger set of websites. The researchers built a system to automatically analyse websites in 24 different languages and their process reviewed privacy policy pages as well as cookie consent notices. Regarding privacy policies, the researchers found that before the GDPR 79.6% of the websites had a privacy policy. After the GDPR that number rose to 84.5% (Figure 2.5a), which is a small increase. However, they found that 85.1% of those websites changed their privacy policy between March 2017 and May 2018 (before GDPR passed and after it was enforced). Interestingly, the researchers determined that on average, the privacy policy text "rose from a mean of 2,145 words in March 2016 to 3,033 words in March 2018". This yields a 41% increase in word count within 2 years as is shown in Figure 2.5b. Furthermore, the researchers found that the adoption of cookie consent notices had increased. More specifically, cookie consent notices increased by 46.1% to 63.2% between January 2018 and October 2018 (Figure 2.5c). Finally, they determined that most websites use a set of 31 cookie consent notices that automatically implement those banners.

Due to all the additional regulation that came into force with the GDPR, one could assume that tracking would be impacted and therefore, websites would track their users less. However, this does not appear to be true and although users have the ability to manage their data better tracking is still ubiquitous on the internet. This was observed by Sørensen and Kosta [27] who investigated whether the GDPR lowered the activity of third-party



(a) Websites offering a privacy policy page before and after the GDPR.

(b) Privacy policy length before and after the GDPR.

(c) Cookie banner implementation before and after the GDPR.

Figure 2.5: The impact of GDPR on the cookie banners and privacy policies of websites, as shown by Degeling et al.

trackers. They did this by measuring the number of trackers before and after the GDPR went into force. The researchers used two main categories for the websites that they chose to survey. Those were the publicly-owned websites (e.g. government) and the privately-owned (e.g. entertainment) ones. The researchers surveyed websites from every country of the European Union. In total, they gathered 1,363 websites from 39 different countries which then were divided into 11 different categories. Firstly, they noticed that public websites store fewer third-party trackers than their private counterparts. However, they did not notice significant changes in the number of Third-Party Cookies, in either category, post-GDPR. Overall, Sørensen and Kosta did not observe a significant decline in the number of third-party trackers after the GDPR was put into force. Thus, the authors noted that the assumption that the GDPR is going to lead in an immediate decline in the number of TPs does not hold.

2.5 Cookie Notices and Tracking

As Sørensen and Kosta showed, online tracking is everywhere even though legislation, such as the GDPR, is trying to contain it and give the ability to users to manage their online privacy. Furthermore, it is very hard for users to know the number of cookies that are being stored on their browsers and therefore the amount of tracking that websites are carrying out.

Unfortunately, users are being tracked by tens, and sometimes hundreds, of Third-Party Cookies (TPs) that are being stored by the websites that they visit and these cookies are usually owned only by a handful of companies such as Google, Facebook and Twitter. The largest survey that looked at the prevalence of online tracking was conducted by Englehardt and Narayanan [1]. They visited 1 million websites in order to detect the occurrence of online tracking on these websites. This included Third-Party Cookies (TPs), cookie synchronisation as well as fingerprinting techniques. In order to do this, the researchers developed OpenWPM, an open-source tool which allows researchers to automatically measure cookies and trackers in websites. OpenWPM allows for stateful measurements of websites. This means that websites don't treat OpenWPM browser instances as new users. This enables measurements of trackers across multiple websites as well as detecting differences in content etc. Furthermore, because of this, OpenWPM allows researchers to load user profiles in order to provide realistic scenarios. OpenWPM automatically stores the results in an SQLite database that is made available to researchers for further analysis. Using OpenWPM, Englehardt and Narayana made over 90-million requests to their 1-million website set. Interestingly, they found that there are 81,000 web trackers on the websites that they surveyed but only 123 of those are the most prevalent. Specifically, they noted that "Google, Facebook, Twitter and AdNexus are the only third-party entities that are present on more than 10% of the websites". Moreover, the researchers found that the adoption of HTTPs technology remained low. In fact, 54% of those trackers are HTTP only. Interestingly, they empirically confirmed the now well-known fact that news websites contain the most trackers compared to other categories within their set. The authors attributed this to the business model of these websites. Moreover, the researchers found that third-parties are "highly connected" by using cookie syncing. More specifically, the top 50 third-parties, which use cookie syncing technology, the probability of finding them in one of the top 100 websites is 66%.

Online tracking using cookies, among other techniques, is rife and oftentimes extremely invasive. However, certain types of websites can employ significantly more tracking than others. For instance, political websites that lean towards the right, appear to have significantly more third-party trackers compared to their left-leaning counterparts.

This was shown by Agarwal et al. [28] who researched whether hyper-partisan websites (HPWs) demonstrate any "particular differential behaviour" when tracking their online users. For example, they wanted to determine whether left-leaning websites track right-leaning users in a different way than the left-leaning ones do. In order to gather accurate data, the researchers curated a number of different personas that correspond to a different demographic. For gathering the data, they used the OpenWPM tool to visit the HPWs. In total, they visited 667 websites with high partisan content. After analysing the data, the researchers showed that right-leaning websites (W^R) set 9 more cooking on average than the left-leaning websites (W^L). Furthermore, they found that 72% of the Third-Party Cookies that all HPWs store are similar to the ones that the "general Web" stores as well. Moreover, the researchers noted that the Alexa ranking of HPW, the W^R always tends to track users more than the W^L. Within these results, the researchers found extreme cases where both W^L and W^R websites store more than 1000 cookies on a user's device. Interestingly, the research showed that the ads shown in W^R are costing 5 times more than the ads on the left-leaning websites. Overall, HPWs that tend to lean on the right, store significantly more Third-Party Cookies compared to their left-leaning counterparts with personas that realistically represent a specific demographic (e.g. Young Man), tend to receive 25% more cookies from HPWs.

2.6 Summary

This section summarised previous research on the topic of cookie banners, privacy policies as well as internet tracking using technologies such as Third-Party Cookies and device fingerprinting. Overall, the work can be summarised in four distinct sections. These are studies on cookie banners and the privacy options that they provide to users, research on whether user location affects those cookie notices, work investigating the effects of legislation such as the GDPR and finally the prevalence of online tracking using cookies.

However, a number of them had a particular influence on this project and especially on its research questions and methodology. Specifically, the work by Habib et al. as well as Nouwens et al. looked at the number as well as the type of privacy options provided to users by websites. This project raises similar questions but aims to do two things differently. Firstly, the objective is to survey a significantly larger dataset from 2 entirely different countries that, to our knowledge, have not been researched before. Secondly, the data gathering, as well as analysis, is conducted in an automated fashion, with manual intervention only when required, as discussed in detail in the next section of this paper. Furthermore, the tools that automate this process can be used by other researchers to conduct similar research in other countries if they wish to.

Englehardt and Narayanan focused on how widespread cookie tracking is on the internet they performed their analysis by creating OpenWPM, which is a powerful privacy measurement tool. The authors have made OpenWPM open and available to the research community. Thus, it has been extended and used by this project to crawl websites and detect cookie banners as discussed. Furthermore, the work by Eijk et al. and their CSS cookie banner scraping extension of OpenWPM has also inspired the scraping method used in this project as well. The next section will discuss in further detail the tools and methodologies used in this project in order to gather as well as analyse data.

3 Methodology

This chapter introduces the methods and techniques used in order to detect and measure cookie banners and the privacy options that they provide. Overall, the methodology of this project can be divided into 4 distinct steps. These steps are summarised, in chronological order, in the following list:

1. **Dataset identification**: Collect a robust set of functioning websites to analyse and extract their cookie banners;
2. **Data collection** Crawl the identified websites and effectively collect the relevant data such as the source code of the cookie notices as well as screenshots of the webpages;
3. **Data sanitisation & normalisation**: Sanitise and structure the collected data from the previous step is a searchable and user friendly data structure.
4. **Data analysis**: Analyse and make sense of the collected cookie banners and their privacy options.

3.1 Data Identification

The first step is to identify websites that can be analysed further in the subsequent steps of this project. Using publicly available lists such as Tranco [29], popular websites for the countries of interest were identified and used for further research. Identifying and utilising the most popular websites, from a wide range of categories, can show how the vast majority of users in a given country experience the web and can illustrate the impact of cookie banners and the privacy options that they offer.

Furthermore, country-specific lists, such as TopGR (`https://topgr.gr`), were also used to identify websites. This was done in order to handle local nuances that would otherwise be missed if this project used only Tranco for building its dataset. More specifically, country-specific lists aided the discovery of popular websites that do not use the Top Level Domain (TLD) of their country of origin. For instance, although British Airways (BA) is a UK company, they are using the ".com" TLD for their website and therefore, it would have been excluded from the dataset if the country-specific lists were not taken into account.

However, it is important to note that not all websites allow crawling and a lot of them explicitly state that they only allow "personal use" of their services and content to their visitors. In order to comply with the terms of use of the websites within the dataset, this project introduces two novel ways of respecting the restrictions imposed by the websites. These are summarised in the following list:

1. A "robots exclusion standard" parser that verifies whether websites allow crawling by reading their robots.txt file;
2. A "terms of service" (TOS) parser that makes best efforts to find exclusionary terms, such as "for personal use only", in order to comply with the TOS of each website.

3.2 Data Collection

The second step is to effectively identify and collect the cookie banners on the websites in the dataset. In order to do so, this project takes advantage of the "I don't care about cookies" (IDCAC) list which provides an extensive selection of common CSS selectors that cookie banners use. More specifically, the CSS selectors from that list are parsed and then added to a database to be used by OpenWPM. Furthermore, an additional 64 selectors are added to the list which was identified during testing.

After the cookie selectors have been parsed and added to the database, OpenWPM uses them to identify the cookie banners within the dataset identified in the previous step. More specifically, OpenWPM has been extended by this project to detect the cookie notices within a given website. OpenWPM checks whether the visited website has any of the selectors from the IDCAC list and if so, the cookie notice's HTML code, as well as its attributes, are added to a database for further analysis. Table 3.1, summarises the data that is being saved in the database after OpenWPM has successfully identified the cookie banner.

Field	Description
HTML	The full HTML code of the cookie banner.
size	The width and height of the cookie banner.
position	The x and y coordinates of the cookie banner in the page.
banner_exists	Indicates whether the banner has been successfully identified by OpenWPM.
selector	The CSS selector that the cookie banner uses.

Table 3.1: The cookie banner attributes that are saved in the database by OpenWPM.

3.3 Data Sanitisation & Normalisation

After OpenWPM has completed crawling the websites in the dataset, the next step is to sanitize and normalize the collected data. To our knowledge, no standard exists for cookie banners and therefore, every website has a different implementation for their notices. Thus, the HTML code, as well as the options that they provide, can be drastically different from website to website. This can make data analysis extremely difficult since there is no clear or efficient way of querying arbitrary data and therefore, the data have to be transformed into a consistent data structure before processing it further.

Firstly, the collected data have to be sanitised so that the privacy options that are offered by the cookie banners are identified and classified. In order to do so, the privacy options are categorised into 4 distinct categories. These are the affirmative, non-affirmative, informative and managerial categories and they are summarised in Table 3.2.

After the data have been sanitised they can be transformed into a consistent data structure that allows for efficient and easy analysis. More specifically, the arbitrary structure of the collected cookie notices is turned into formal SQL-like table format. This is done for the following reasons:

1. **Accessibility**: Not everyone understands HTML and therefore, it might be hard for someone without programming skills to analyse the gathered data. Furthermore, not all websites use simple HTML for their cookie banners, like the one seen in Figure 3.1a.

2. **Efficiency**: Even experienced HTML developers might have a hard

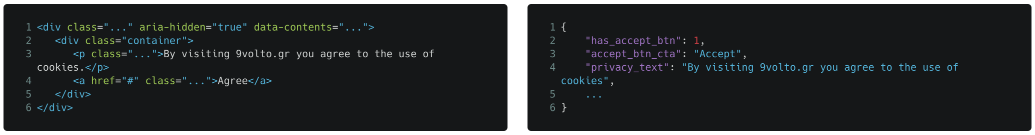
Category	Description
Affirmative	Options that prompt users to accept the use of cookies such as “Accept all”.
Non-Affirmative	Options that allow users to opt-out from cookie tracking such as “Decline”.
Informative	Options that take users to informational pages such as the Privacy Policy page.
Managerial	Options that allow users to manage tracking options and opt-out from certain trackers.

Table 3.2: The categories that the privacy options are classified into.

time analysing the thousands of cookie banners that were collected during this project.

3. **Searchability**: Performing queries can be inefficient and inaccurate when having to search for this kind of data structure.

For instance, Figure 3.1a, depicts the HTML code from the HTML code of a cookie banner found in 9volto (www.9volto.gr) which contains a sentence informing users that the website uses cookies and an “Agree” button. After the sanitisation and normalisation steps, discussed in this section, the cookie banner is transformed into the data structure shown in Figure 3.1b. Therefore, this can then be transferred to a MySQL table or an Excel sheet for efficient and easy analysis. The complete SQL-like data structure is summarised in Table A.1.



```html
<div class="..." aria-hidden="true" data-contents="...">
   <div class="container">
      <p class="...">By visiting 9volto.gr you agree to the use of cookies.</p>
      <a href="#" class="...">Agree</a>
   </div>
</div>
```

(a) HTML code from 9volto.gr.

```json
{
    "has_accept_btn": 1,
    "accept_btn_cta": "Accept",
    "privacy_text": "By visiting 9volto.gr you agree to the use of cookies",
    ...
}
```

(b) Normalised cookie banner.

Figure 3.1: HTML code from a cookie banner before and after data sanitisation & normalisation.

3.4 Data Analysis

After the data has been sanitised and normalised in a table-like data structure, the next step is to analyse the data in order to answer the research questions. While the data can be exported in a number of different applications, such as Microsoft's Excel, this project retains the data in the database that was created in the previous steps in order to take advantage of SQL's powerful and fast querying syntax and features. Where SQL fails to extract the required data, Python scripts have been used to perform more advanced analysis.

As before, the results are presented in a table-like format that can be converted into useful plots or exported in other applications for further analysis.

3.5 Summary

This chapter introduced the methods used and developed to detect cookie banners and analyse the privacy options that they give to users. In summary, this project's methodology consists of 3 distinct steps. Firstly, popular websites from the target country are identified using open-source lists such as Tranco. Importantly, this project takes extra steps and makes best efforts to comply with each website's rules on crawling. Secondly, the cookie banners and their privacy options are identified and saved with the aid of OpenWPM as well as building on previous similar research to improve performance as well as retrieve more accurate data. The third step introduces a novel technique that structures the collected data in table format to allow for easy and efficient analysis. Finally, the structured data are analysed using SQL and Python in order to answer the research questions set by this project. Figure 3.2, depicts the three methodology steps, as well as their sub-tasks, that were discussed in this chapter.

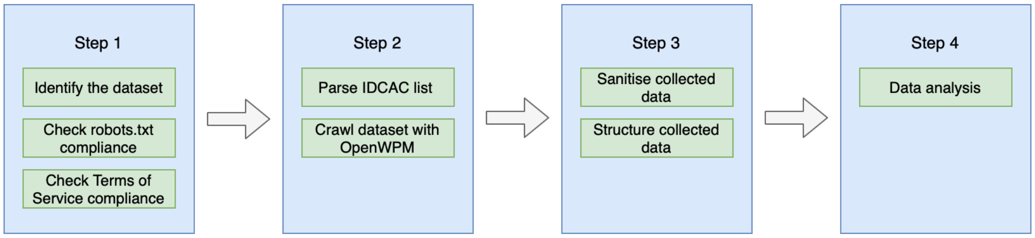


Figure 3.2: The 4 steps of this project and their sub-tasks.

4 Implementation

This chapter discusses the implementation details of the techniques used to crawl, extract and analyse the cookie banners in Greece as well as the UK. The implementation steps can be divided into 4 distinct steps which are summarised in the following list:

1. **Identifying websites**: The methods and software developed to build a robust dataset of popular websites both in Greece and the UK and to deal with the issues faced from "local nuances";
2. **Collecting the cookie banners**: The extension that was developed to allow OpenWPM to detect and collect cookie banners on the websites that it visits;
3. **Normalising the data**: The methods and software used to identify the privacy options within the collected cookie banners and how they were classified based on the categories introduced in Table 3.2;
4. **Analysing data**: The techniques used to extract information from the collected data in order to answer the research questions set out by this project.

4.1 Identifying Websites

The first step of the cookie banner collection process is to identify the most popular websites within Greece and the UK. However, while it might be trivial to single out websites that use each country's TLD, identifying popular websites with different suffixes can be challenging. Furthermore, some websites explicitly refuse tracking or crawlers in their website and therefore, they have to be excluded from the dataset. Specifically, the first step can be subdivided into 3 sub-steps that are summarised in the following list:

1. **Gathering the URLs**: The global, as well as local, ranking lists used and the software, developed to extract the links from those lists;

2. **Compliance with robots.txt**: The novel method built and used by this project in order to comply with the Robots Exclusion Standard set by each website;

3. **Compliance with Terms of Service**: The novel method developed by this project in order to make best efforts to comply with the terms and conditions of each website in the dataset.

4.1.1 Gathering links

The primary focus of this project is to study cookie notices in Greek as well as UK websites. Therefore, it is essential to identify websites in both countries and ideally the most popular ones there in order to build a robust and diverse dataset.

For identifying the top websites in the target countries, this project utilised the Tranco list which was created by Le Pochat et al. Tranco, aggregates the top results from lists provided by Alexa (`https://alexa.com`), Cisco Umbrella (`https://umbrella.cisco.com/`), Majestic (`https://majestic.com/`), and Quantcast (`https://quantcast.com/`) and makes them available in a CSV format. In total, the Tranco list contains 1-million websites from across the world ordered by their rank.

Filtering websites by their TLDs & SLDs

Both Greece and the UK have their own Top-Level Domain (TLD) which are .gr and .uk respectively. While Greece was also granted the "el" (or ".ελ" in Greek) TLD by IANA, after petitioning for years [30], it has seen a low adoption by businesses and government agencies and therefore, it has not been taken into consideration in this project. Therefore, since each TLD is unique to a country, they can be used to find and retrieve websites from Tranco's large dataset.

Furthermore, Second-Level Domains (SLDs), such as .gov.uk, can also be used to enhance the filtering and also demonstrate the category that a website falls into. Thus, this project takes advantage of a wide range of Top and Second-Level Domains, in order to identify the websites in Greece as well as the UK. Table 4.1, summarises the TLDs and SLDs that are used to filter and retrieve websites from the Tranco list.

Greece	UK	Description
.gr	.uk	The Top-Level Domain for each country.
.com.gr	.co.uk	For private businesses and the commercial sector
.gov.gr	.gov.uk	Used by local and federal government agencies
.net.gr	.net.uk	Used by Internet Service Providers (ISPs)
.org.gr	.org.uk	Primarily used by non-profit organisations
sch.gr	.sch.uk	For educational authorities such as high schools
	.ac.uk	Used by higher education institutes
	.{ltd/plc}.uk	For Ltd or Plc companies
	.me.uk	Personal names
	.nhs.uk	Used by the NHS and its trusts
	.police.uk	For the UK police and its local forces

Table 4.1: The TLDs and SLDs for Greece [2] and the UK [3,4] used to filter the Tranco list.

Dealing with local nuances

However, it is not mandatory for businesses and other websites to use the conventional .gr or .uk suffix. Therefore, popular websites from each country that are using TLDs such as .com are missed when using the TLD as the only filter.

For instance, Aegean Airlines (`https://en.aegeanair.com/`) is the largest Greek airline, providing domestic, as well as international flights, to and from a large number of airports across the world. Although it is a Greek business, since they are operating globally they are using the popular .com suffix to provide a more familiar domain name to international customers and therefore, attract more business. Thus, although Aegean is in the Tranco list, due to the fact that they use a .com domain will be missed if a simple TLD/SLD clarification is used.

For tackling the local nuances discussed above, additional lists have been used to aid the original Tranco list. These lists rank the most popular websites within a certain country without taking into consideration their TLD. More specifically, these lists are:

- **TopGR** (`https://topgr.gr`): Provides a list of the most visited Greek websites, based on their Alexa rankings;
- **Kadaza** (`https://kadaza.co.uk`): Provides the 24 most popular websites in the UK in a wide range of categories such as news, shop-

ping and travel. The list is manually curated by the Kadaza team by "monitoring the latest website traffic data and local website trends";

- **Finder.com** (`https://finder.com`): Compares online online retailers and services for a number of different categories such as beauty, fitness, energy providers, mortgages etc. Finder.com manually curates its lists and updates them daily in order to ensure best quality. Because of the rise of online shopping due to the Coronavirus pandemic [31, 32], ensuring that all major retailers are included in this study is essential. Therefore, Finder provided a reliable source for finding UK online retailers.

While the above lists have not been used as the primary source for building the dataset, they provide a good solution on the issue of local nuances and they also provide a more diverse set of websites that may have never been studied before.

Building the dataset

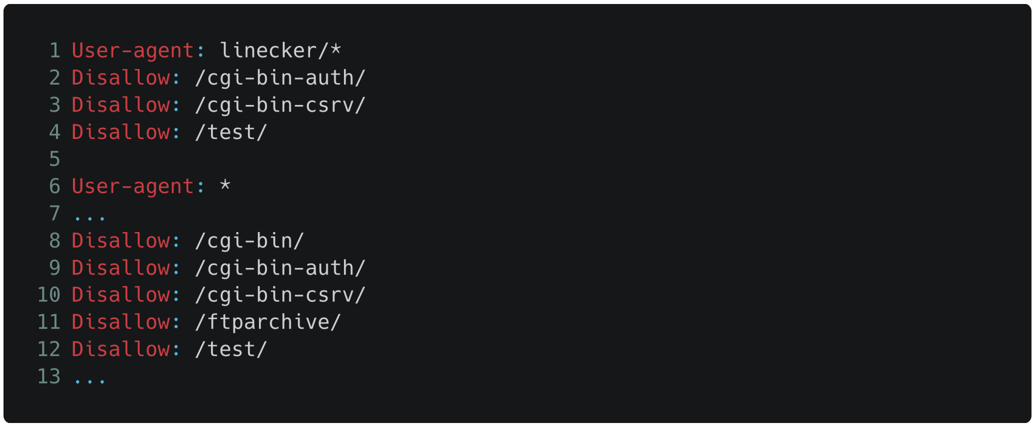
In order to build the dataset, a Python3 script was developed that applied the TLD/SLD filter on the Tranco list and also parsed the additional lists. The script takes 3 steps to build the dataset, which is summarised in the following list:

1. **Build the TLD set**: The script parses the Tranco list and builds a set that contains all the websites with the specific suffix (e.g: .gr);
2. **Build the local set**: The script parses the additional lists (e.g. TopGR) and builds a set containing the websites from there;
3. **Merge the two sets**: The script merges the two sets. The final set contains the original set (e.g all the .gr domains) and the items of the second set that also exist in the Tranco list.

Finally, the remaining websites are stored in an SQLite database in order to be easily accessible for further analysis. SQLite was chosen for the purposes of this project since it provides a fast and reliable SQL database implementation. However, the most important feature of SQLite is that it is self-contained and therefore, the database file can be easily shared and accessed, without the need for an SQL server. The source code for the above script can be found in Appendix A.2.1.

4.1.2 Compliance with the Robots Exclusion Standard

The Robots Exclusion Standard, or simply known as robots.txt, can be used by websites to specify which parts of the website should not be processed by web crawlers and robots [33]. Robots are used by search engines for categorisation purposes but can also be used by competitors to retrieve information about the website. Therefore, website administrators can choose to exclude certain directories from crawlers as shown in Figure 4.1.



```
User-agent: linecker/*
Disallow: /cgi-bin-auth/
Disallow: /cgi-bin-csrv/
Disallow: /test/

User-agent: *
...
Disallow: /cgi-bin/
Disallow: /cgi-bin-auth/
Disallow: /cgi-bin-csrv/
Disallow: /ftparchive/
Disallow: /test/
...
```

Figure 4.1: Parts from the York University robots.txt file (`https://york.ac.uk/robots.txt`) that uses the Disallow keyword to stop crawlers from processing that part of the website.

Since the software built and used for detecting cookie banners resembles a crawler, this project aimed to respect the exclusion standard set out by the websites. To do so, a Python3 script was built that parses robots.txt files and verifies whether they allow crawling. That script was applied to the dataset that was built on the first step.

More specifically, the script verifies that a robots.txt exists and it allows crawlers to parse the root directory of the website, otherwise the website is marked as "not-crawlable" as it can be seen in Listing 4.1.

```
For every $website in the dataset:
    Find the robots.txt file of $website
    If robots.txt does not exist:
        Mark $website "uncrawlable"

    If robots.txt allows crawling:
        Mark $website "crawlable"
    Else:
        Mark $website "uncrawlable"
```

Listing 4.1: Pseudocode of the algorithm followed by the robots.txt parser script.

The script utilises the reppy Python library [34] that aids with parsing and analysing the robots.txt of a website. Furthermore, it has been parallelised using Python's multiprocessing library in order to take advantage of every available CPU core in the system and reduce runtimes. Finally, the SQLite database containing the dataset is updated to reflect the crawl status of each website. The source code for the robots.txt compliance script can be found in Appendix A.2.2.

4.1.3 Compliance with the Terms of Service

In addition to the Robots Exclusion Standard, this project makes best efforts to comply with the Terms of Service (TOS) of the websites within the dataset. This was done for two reasons:

1. Some websites use default or auto-generated robots.txt file that allows crawlers to visit their website. However, on their TOS they specifically state that they do not want robots scraping their content;
2. The cookie banner detector software developed for this project collects the HTML of the cookie notices found in websites, which can be considered content. Some websites explicitly state that their content is aimed for personal use only.

The exclusionary terms were selected after manually reading over 50 TOSs, from websites in both countries, and finding the most common wording that websites use for prohibiting visitors from using the website, as well as its content, in a not-acceptable manner. A number of different terms and phrases both in English as well as Greek were identified, such as "for personal use only" which are summarised in Table 4.2.

A Python3 script was developed that makes best efforts to ensure TOS compliance. More specifically, the script visits a website and looks for a TOS link. If a link is found, it visits the Terms of Service page and examines its text for terms or phrases that may indicate that the website refuses to be tracked. The algorithm used by the script is shown in Listing 4.2.

```
For every $website in dataset And robots.txt allows crawling:
    If $website has TOS link:
```

Greek Phrases	English Phrases
Strictly for personal use	Personal use only
Only for personal use	Only for personal use

Table 4.2: The English as well as Greek phrases (translated) that the Terms of Service parser use. The original Greek terms can be found in A.2.

```
        Go to the TOS page
        If TOS page has exclusionary terms:
            Mark website "uncrawlable"
        Else:
            Mark website "crawlable"
    Else:
        Mark website "crawlable"
```

Listing 4.2: Pseudocode of the algorithm that the Terms of Service parser uses.

The Terms of Service parser uses Python's "request" library to navigate to websites and BeautifulSoup4 [35], a library that provides easy and efficient programmatic access to HTML elements. More specifically, the steps are taken by the TOS parser are summarised in the following list:

1. **Visit the website**: Using Python's request library, the script makes a simple GET request to the URL and it pulls its HTML code for analysis;

2. **Look for a TOS link**: The HTML code is then parsed by BeautifulSoup. The script looks for anchor elements (¡a¿) that might link to a Terms of Service page by searching for specific terms which are listed in Table A.3;

3. **Go to TOS page**: If such a link is found, then the script pulls the HTML code from the TOS page and it feeds it to the HTML parser as before;

4. **Look for exclusionary terms**: The script searches the Privacy Policy text for specific exclusionary terms from Table 4.2.

Finally, the script updates the crawl status of each website in the dataset. The full source code of the Terms of Service parser can be found in Appendix A.2.3.

4.2 Detecting Cookie Banners

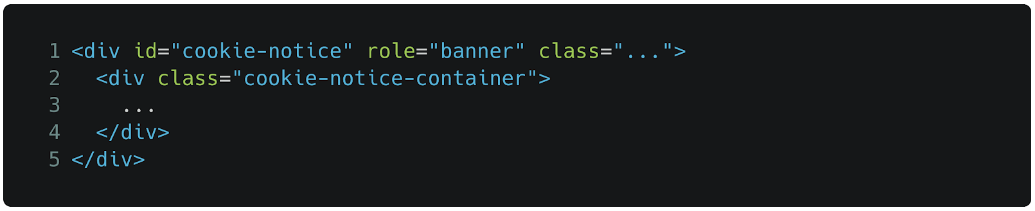
After the dataset has been built, the next step is to visit the candidate websites that allow crawling. In order to do so, this project uses OpenWPM and extends it by giving it cookie banner detection capabilities. More specifically, detecting cookie banners can be split into two sub-steps which can be summarised as follows:

1. **CSS Selectors**: Build a list of common CSS selectors that are used by cookie banners across Greece and the UK;
2. **Extend OpenWPM**: Give open OpenWPM cookie banner detection capabilities and scrape the dataset with the aid of the CSS Selectors list that was built before;
3. **Data collection**: Collect the cookie banners and their attributes for later analysis.

4.2.1 I don't care about cookies

The GDPR requires websites to seek the users' permission before installing tracking cookies on their web browser. However, if users surf the web anonymously then the same websites will ask for the user's permission again. "I don't care about cookies" (IDCAC) is a popular web browser extension that removes the cookie notices and saves the user "thousands of unnecessary clicks".

In order to remove those notices, IDCAC has built a list of CSS selectors that websites use for their cookie notices. For example, Figure 4.2 depicts the HTML of a cookie banner found in SkyExpress, a popular Greek airline (`http://skyexpress.gr`). Notice that it has an "id" attribute which is set to "cookie-notice". Thus, if the IDCAC list contains the CSS selector `#cookie-notice`, the extension can effectively hide the cookie banner.



```html
<div id="cookie-notice" role="banner" class="...">
  <div class="cookie-notice-container">
    ...
  </div>
</div>
```

Figure 4.2: The HTML of the cookie banner found in SkyExpress.

A similar technique was used by Eijk et al. when studying the cookie banners on their dataset. However, the implementation of the IDCAC parser developed for this project differs significantly. The following list summarises the steps that are taken to parse the IDCAC as well as their differences between this project and Eijk et al. approach:

1. **Download**: Retrieve a fresh copy of the "I don't care about cookies" list from their website (`https://www.i-dont-care-about-cookies.eu/abp/`), to ensure that no new CSS selectors are missed. While Eijk et al. download and parse the IDCAC list every time their code is executed, this project processes the IDCAC list in a separate script (or a different sub-step) and adds the CSS selectors in a database. This was done for 2 reasons. Firstly, the IDCAC list is not updated regularly and therefore, there is no point in downloading it and parsing it every time the code is executed. Secondly, the set of websites parsed by OpenWPM are almost always the same and therefore, it is unlikely that OpenWPM will have to deal with different CSS selectors and cookie notices at every run;

2. **Parse**: Remove any unnecessary lines, comments or syntax around the CSS selectors and keep only the required content from the list. The IDCAC cookie parser implemented for this project is significantly optimised for better performance compared to the one introduced by Eijk et al. For instance, they use Python's split method [36] to process different parts of the IDCAC list which can have a worst-case performance of $O(n)$. On the other hand, this project's parser takes advantage of Python's string cutting and slicing, which has a performance of $O(1)$. This can have a significant impact when processing over $16,000$ CSS elements;

3. **Save**: Store the CSS selectors in a database for caching purposes, as well as easy access, by other applications such as OpenWPM, as discussed above.

The IDCAC about cookies list has been enriched with an additional 64 CSS cookie banner selectors that were identified while testing the code of this project. While these 64 CSS selectors only account for 0.4% of the IDCAC list, without them at least 400 Greek websites would have been missed, including online retailers and government agencies. These additional CSS selectors have been submitted to the IDCAC project in order to be added to the final list and improve IDCA's reach in Greece. The full source code of the IDCAC parser and the additional CSS selectors can be found in Appendix A.2.4.

4.2.2 Making OpenWPM care about cookies

OpenWPM is an open-source tool, developed by Englehardt and Narayanan. It allows researchers to automatically measure Third-Party Cookies, trackers, cookie synchronisation as well as fingerprinting techniques, in websites.

Unfortunately, OpenWPM has no way of detecting the existence of cookie notices on the websites that it visited. Therefore, this project has developed a novel way of extending the functionality of OpenWPM in order to effectively detect cookie banners and store relevant information about them.

More specifically, for each website OpenWPM checks whether it contains a CSS selector from the cached IDCAC list, using Selenium [37], which allows for searching the HTML code of the website. Once a selector has been identified, additional checks are made to ensure that this is not a false-positive. These additional checks are summarised in the following list:

1. **Selenium check**: Ensure that Selenium returned valid HTML e.g. the length of the HTML;

2. **Cookie check**: Verify that the HTML returned by Selenium contains the terms “cookie” or “cookies”.

If the cookie banner has been successfully identified, then it is stored in a database for further analysis. The combination of the cached cookie selectors and Selenium’s efficient DOM [38] search enables the cookie banner extension to be fast, efficient and fault-tolerant.

Listing 4.3, summarises the novel steps added to OpenWPM in order to detect the cookie banners.

```
For every $website in dataset:
    Visit $Website
    If $website has CSS selector:
        Get cookie banner HTML
        Ensure not false-positive
        Save cookie banner
        Move to the next website
```

Listing 4.3: Pseudocode of the steps that the OpenWPM extension uses.

Furthermore, the cookie banner extension developed for this project is vastly different from the one developed by Eijk et al. for the purposes of their study. Specifically, the differences between the two extensions are summarised in the following list:

1. **Searching for a cookie banner**: Eijk et al. search each website for every CSS selector on their list. Thus, if they have n CSS selectors, they perform n iterations and DOM searches (which can also be expensive) on each website, even if the correct CSS selector has already been found. This was probably done to eliminate false-positives (e.g. a selector matched an element that is not a cookie banner). On the contrary, the cookie banner detection implementation for this project stops when the first CSS selector matches an element, potentially saving thousands of unnecessary loops per website. During testing, and also when running the experiment on the full set of websites, the number of false-positives was not significant enough to justify the performance hit of an exhaustive search and those "edge" cases were fixed manually;

2. **Data output**: This project outputs significantly more information and data for later analysis compared to Eijk et al. In addition to the cookie banners' source code, the full website source code, as well as screenshots, are saved which can help when analysing the cookie banners.

In total, the cookie banner extension consists of 7 files across the existing OpenWPM project. Table 4.3, summarises these files, their purpose and source code.

File	**Description**	**Code**
CookieBanner.py	Cookie banner representation	A.2.10
CommandSequence.py	OpenWPM API to allow users to call cookie banner extensions	A.2.8
Types.py	The types of commands	A.2.11
browser_commands.py	The cookie banner detection implementation	A.2.6
command_executor.py	Works with CommandSequence to call the right methods	A.2.7
cookie_utils.py	Helper methods to aid cookie banner detection	A.2.9
openwpm_cookie_parser.py	A script that prepares and starts OpenWPM	A.2.5

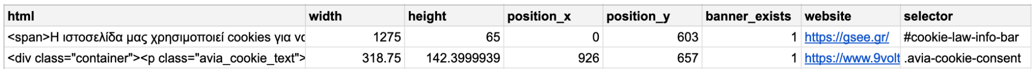
Table 4.3: The files extended to allow OpenWPM to detect cookie banners.

4.2.3 Data Gathering

In addition to the cookie banner detection capabilities, the data gathering capabilities of OpenWPM were also enhanced as part of this project. More specifically, after a cookie banner is detected, a number of its attributes, as well as its source code, are stored in a database for further analysis. The information gathered per cookie banners include:

- A boolean value that indicates whether a cookie banner was found for a given website;
- The size and position of the cookie banner;
- The HTML source code of the cookie banner;
- The CSS selector that was used to identify the notice,

Figure 4.3, shows sample data collected by two websites using the cookie banner extension. Moreover, a screenshot, as well as the entire HTML code, is saved for every website. This is done in order to manually inspect errors or false-positives that may occur during the runtime of OpenWPM.



html	width	height	position_x	position_y	banner_exists	website	selector
<span>Η ιστοσελίδα μας χρησιμοποιεί cookies για vc	1275	65	0	603	1	https://gsee.gr/	#cookie-law-info-bar
<div class="container"><p class="avia_cookie_text">	318.75	142.3999939	926	657	1	https://www.9volt	.avia-cookie-consent

Figure 4.3: Sample data gathered after OpenWPM has finished running.

4.3 Sanitising & Normalising the Data

Throughout its execution, the OpenWPM cookie banner detector stores a plethora of data. However, given the amount of data and their particular structure, it can be challenging to efficiently query and analyse them. To tackle those issues, the following steps have been taken in order to make the data easier and more efficient to search:

1. **Identify actions**: Determine the “call to action” words used in the options offered by the cookie banners. For instance, distinguish the affirmative options such as “Accept” or “Ok” from the negative options such as “Decline” or “No”.

2. **Structure data**: Having identified the options provided by the cookie banners, the collected data from the previous step can be converted from an arbitrary HTML format to a structured table-like format which allows for easier querying.

4.3.1 Identifying Actions

The first step of normalising the data is to determine the options that cookie banners provide and their different variations. The cookie banner options categories are the affirmative, non-affirmative, informative and managerial as they were introduced in Table 3.2.

In order to categorise the privacy options from the collected data, a Python3 script was developed. More specifically, the script looked at the HTML of each individual cookie banner, stripped it from all the unnecessary HTML tags and kept only the `<a>` and `<button>` tags. This was done because the overwhelming majority of privacy options (almost 100% in the dataset) are programmed using those tags for 2 reasons. Firstly, they are easily identifiable by the users and know that they can interact with them. Secondly, it is easy for developers to detect when users have clicked such an element and run additional code. As a final step, the text within those tags, which is also known as “call to action”, was stored in a simple SQL table.

After the above script finished running, the SQL table and its entries were manually inspected and classified into the 4 cookie banner option categories. While this project aims to automate the entire process of detecting and classifying cookie banners, manual inspection of the privacy options and their call to actions can be beneficial and in some cases unavoidable.

Similarly to when identifying popular websites for a country, call to action phrases can hide local nuances. For instance, the majority of cookie banners collected from Greece had the noun “αποδοχή” (I accept) as their affirmative call to action. However, a large number of them used the verb “δέχομαι” which also means “I accept” as an affirmative call to action.

It is clear that two different words can have the exact same meaning and therefore, a generic program or someone without an understanding of the language and its nuances can easily fail to classify the privacy options properly. Thus, manual intervention and classification, instead of an automated program, was decided for this step in order to avoid missing valuable information from an already rich dataset.

4.3.2 Structuring the Data

After all the phrases and call to actions have been properly categorised, the arbitrary data collected by OpenWPM can be converted into an easy to search data structure. More specifically, the goal of this step is to take the HTML code collected from the cookie banners of each website and turn it into a SQL-like table that allows for efficient queries, without losing the core information included in the cookie banners. Converting the inconsistent structure of HTML into a structured format can yield the following benefits when analysing the data:

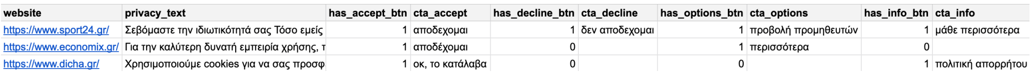
1. **Accessibility**: Knowledge of HTML is not required when querying the data and therefore, the data can be imported in more popular applications such as Microsoft Excel;
2. **Efficiency**: Faster interpretation and understanding of the data as they are displayed in a tabular format with consistent fields instead of arbitrary HTML code;
3. **Searchability**: Efficient queries using tools such as SQL or Excel's queries, instead of having to programmatically parse HTML code, which can be extremely inefficient and inconsistent.

In order to convert the data to the table-like structure discussed above, a Python3 script was developed and its source can be found in Appendix A.2.12. More specifically, the script uses BeautifulSoup's API to extract every privacy option from the cookie banner and then categorised them accordingly. Once a category has been identified the call to action is stored in the database and the cookie banner is marked to indicate that it contains that category. For instance, if a cookie banner has a button with a call to action "Accept", then its database entry shows that it has an affirmative option. The algorithm described here is shown in Listing 4.4.

```
For every $cookie_banner in the database:
    For every $privacy_option in $cookie_banner:
        If $privacy_option is affirmative:
            Set $cookie_banner has affirmative option
            Save call to action
        If privacy_option is non-affirmative:
            Set $cookie_banner has non-affirmative option
            Save call to action
        ... (Same for informative and managerial categories)
```

Listing 4.4: Pseudocode of the algorithm followed by the Cookie Banner Options parser.

Finally, the script also saves the privacy text that is displayed on the cookie banner, using a similar technique as above. More specifically, using BeautifulSoup it strips the cookie banner from all the unnecessary HTML tags (e.g. <a>, <button>, <img> etc.) keeping only the ones that are used for text such as <p>, <span> etc. Figure 4.4, depicts the table-like format that is produced after the script has finished executing.



website	privacy_text	has_accept_btn	cta_accept	has_decline_btn	cta_decline	has_options_btn	cta_options	has_info_btn	cta_info
https://www.sport24.gr/	Σεβόμαστε την ιδιωτικότητά σας Τόσο εμείς	1	αποδεχομαι	1	δεν αποδεχομαι	1	προβολή προμηθευτών	1	μάθε περισσότερα
https://www.economix.gr/	Για την καλύτερη δυνατή εμπειρία χρήσης, τ	1	αποδέχομαι	0		1	περισσότερα	0	
https://www.dicha.gr/	Χρησιμοποιούμε cookies για να σας προσφ	1	οκ, το κατάλαβα	0		0		1	πολιτική απορρήτου

Figure 4.4: Normalised data after the Privacy Options parser has finished running.

4.4 Querying the Data

After the data has been normalised, it can then be searched in order to answer the research questions. While the normalised data can be extracted to a spreadsheet and open by an application such as Microsoft Excel, this project retained the data in the original database in order to take advantage of SQL's advanced search capabilities. The following list summarises the types of queries that have been developed to retrieve the required data:

1. **SQL queries**: Standard SQL queries, such as SELECT, that allow easy and efficient querying of the data in the database;

2. **Python queries**: Where SQL fails to provide adequate representation of the data, Python has been used to compensate for those queries.

4.4.1 Querying with SQL & Python

The primary goal of Step 3 was to structure the collected data into an easy-to-query data structure. This can be achieved by using standard SQL syntax [39] and most research questions have been answered using that.

However, there are instances where SQL cannot handle the types of calculations required to answer some of the research questions. Those instances have been handled by implementing the queries using Python. For example, Questions 2 and 3 required a more advanced representation of the data and it would have been challenging to achieve those results using SQL and therefore, Python was used instead. Table 4.4, summarises the scripts that were developed to answer the research questions and which language they are written in.

Research Question	**Language**	**Source Code**
RQ1	SQL	A.2.13
RQ2	Python	A.2.14
RQ3	Python	A.2.14
RQ4	SQL	A.2.15
RQ6	SQL	A.2.15
RQ5	SQL	A.2.16
RQ7	Python	A.2.17

Table 4.4: The research question queries and the programming language used to develop them.

4.4.2 Determining the Term Frequency

In order to determine the most common terms in the privacy text of the cookie banners (RQ7), the Term Frequency Inverse Document Frequency (TF-IDF) method has been used [40, 41]. With TF-IDF every term is weighted by dividing its frequency by the number of documents in the corpus, instead of representing that term by its raw frequency. Here, documents refer to the cookie banner privacy text in the dataset.

The first step is to calculate the Term Frequency (TF) of every term in the privacy policies dataset. This is done by dividing the number of occurrences of that word by the total number of words in the document as shown equation 4.1:

$$t, f_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \tag{4.1}$$

The next step is to calculate the Inverse Data Frequency (IDF) of the privacy policy text dataset. This is the log of total documents divided by the number of documents that contain the word was shown in Equation 4.2. Inverse Data Frequency is used for determining the weight of rare words across every cookie banner in the dataset.

$$idf(w) = log(\frac{N}{df_t}) \tag{4.2}$$

The third and final step is to calculate the TF-IDF of the corpus, which is simply the TF multiplied by the IDF as shown in Equation 4.3:

$$w_{i,j} = t, f_{i,j} \times log(\frac{N}{df_t}) \tag{4.3}$$

For simplicity, the average of the TF-IDF of every privacy term is calculated and the 50 most frequent terms are included in the final output. The TF-IDF for this project was implemented in Python3 and the source code can be found in Appendix A.2.17.

4.5 Summary

In conclusion, this chapter discussed in detail the techniques that this project employed to crawl, detect and collect the cookie banners of websites in Greece and the UK alike. Specifically, there are 4 distinct implementation steps.

Firstly, the dataset has to be built and consist of a rich and diverse set of websites which is collected by open-source lists such as Tranco. Furthermore, best efforts are made to comply with the Robots Exclusion Standard and Terms of Service of each website. Then, OpenWPM has to be extended and "learn" how to identify cookie banners in websites and then collect them. This is achieved with the aid of "I don't care about cookies", an open-source CSS selectors list. Thirdly, the arbitrary data collected by OpenWPM have to be transformed in a data structure which allows for easy and efficient analysis. This is done after the privacy options provided by the cookie banners are manually inspected and categorised. Finally, after the data have been structured, they can be further analysed using SQL as well as Python. Figure 4.5, depicts the 4 implementation steps discussed in this chapter.

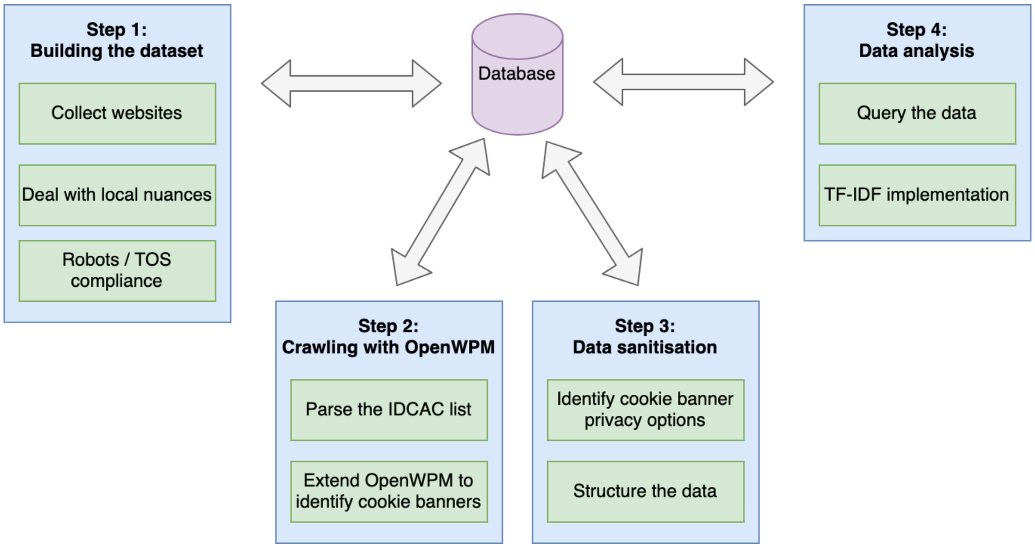


Figure 4.5: The 4 implementation steps and their sub-tasks.

5 Data Analysis

The aim of this chapter is to summarise the size of the dataset that was built prior, as well as after the end of this project. Furthermore, it will discuss the computing resources required to undertake a large-scale cookie banner collection crawl. To follow, a detailed discussion of the results yielded after the data analysis of the collected data was conducted. The following list summarises the core topics that this chapter covers:

1. **Dataset size**: To summarise the amount of data that was collected before and after the crawl took place;
2. **Data analysis results**: To introduce and discuss in detail the results from the data analysis step.

5.1 Collected Data

Before being able to run OpenWPM and the cookie banner detection extension developed for this project, the candidate website dataset had to be built using Tranco, as well as other lists. Furthermore, during the crawl, OpenWPM independently collects data and information about the visited websites such as the HTTP Redirects, Responses and Third-Party Cookies stored in the user's browser.

While the methods and implementation of the cookie banner collection step have been discussed in detail no mention has been made about the actual size of the dataset or the computing resources required to complete the crawl. This section aims to summarise all the above.

5.1.1 Collected Websites

The Tranco list contains a total of 1,000,000 websites. From there, 3,446 are .gr websites and 18,768 are .uk ones. The additional lists used to deal with the local nuances provided an additional 674 websites in which 40 were for the Greek dataset and 634 were for the UK one. Furthermore, 125 websites were removed from the dataset as they were not working anymore. In total, the dataset contained 22,458 websites in which 3,361 were Greek websites and 19,097 UK ones.

Each website was checked in order to determine whether they allow crawlers. The robots exclusion standard parser yielded that 3,157 Greek (93%) and 15,410 UK (69%) websites allowed crawling. Then, after parsing the Terms of Service for each website, it was found that 3,087 Greek (91%) and 14,650 UK (65%) would permit this project to crawl them. Table 5.1, summarises the websites in the datasets that allowed to be crawled.

	Greece	UK
Websites	3,361	19,097
Robots.txt	-204	-3,687
Terms of service	-70	-760
Total	**3,087**	**14,650**

Table 5.1: Total number of websites per country and how many were excluded due to the robots.txt and TOS compliance.

5.1.2 Data from OpenWPM

OpenWPM played an integral role in successfully crawling websites and collecting their cookie banners. During a crawl, apart from the cookie banners, OpenWPM collects a plethora of data for each website that it visits. This includes information about the HTTP Requests and Responses, the scripts that a website loads as well as the number and type of the cookies that are stored in a user's web browser.

In total OpenWPM collected more than 15-million datapoints during the crawls in Greece and the UK. Table 5.2, summarises the total number of datapoints as well as their types.

5.1.3 Computing Resources

Crawling and collecting the cookie banners for thousands of websites can be a highly computing-intensive task. Even though the software developed for the requirements of this project was heavily parallelised, a standard laptop

	Greece	UK	Total
Callstacks	245,874	811,558	1,057,432
Crawl history	11,997	42,830	54,827
HTTP Redirects	25,467	101,362	126,829
HTTP Requests	452,137	1,422,078	1,874,215
HTTP Responses	479,160	1,465,687	1,944,847
Javascript	1,530,712	6,157,821	7,688,533
Cookies	102,378	2,270,009	2,372,387
Navigations	26,615	85,670	112,285
Site visits	3,087	14,650	14,650
Total	**2,877,427**	**12,371,665**	**15,249,092**

Table 5.2: The total number of datapoints collected independently by OpenWPM during the Greek & UK crawls.

(Macbook Pro, 1GB RAM, 2.2 GHz Quad-Core Intel Core i7) would still require over 24 hours in order to complete the crawl for Greece. Similarly, since the UK dataset contained 4 times more websites than the Greek one, the crawl would also take 4 times longer on the same laptop. This can impact testing, data accuracy and most importantly, repeatability of the project.

In order to overcome the above limitations, University of York's "Viking" [42] cluster was used. Viking is a high-performance computing cluster that consists of 173 nodes with a total 42TB of memory and 7024 Intel cores. While only a fraction Viking's resources were used (128GB of memory and 32 cores), the runtime of the crawls was reduced significantly. More specifically, the experiment took a little over 36 hours for the UK and approximately 8 hours for Greece, yielding a substantial performance improvement.

5.2 Results

At the end of the crawl, the data were normalised and then analysed using the methods and techniques shown in Table 4.4. This section summarises the results yielded by the above scripts for both countries surveyed as part of this project.

5.2.1 Prevalence of Cookie Banners

After the data were normalised, the first query performed to the dataset was to determine the number of cookie banners detected by OpenWPM. Specifically, there were 1,497 websites (49%) with cookie banners in Greece and 6,413 websites (53%) in the UK. The cookie banner prevalence in these two countries is summarised in Table 5.3.

	Greece	UK	Greece + UK
Total websites	3,031	11,930	14,961
Cookie Banners	1,497	6,413	7,910
%	**49.3**	**53.7**	**52.8**

Table 5.3: The prevalence of cookie banners in Greek & UK websites.

However, the above results only show the amount of cookie banners that were detected during the crawl and do not indicate the number of websites that store Third-Party Cookies (TPs) in the user's browser. In order to determine this, the data collected automatically by OpenWPM were used. Specifically, they showed that in Greece a total of 1,871 websites (61%) stored at least one TP. Similarly, the number of UK websites that stored a minimum of one TP in the user's browser was 8,256 (70%).

Therefore, it is obvious that 374 Greek (12%) and 1843 UK (15%) websites have not implemented a cookie notice even if they utilise third-party cookie tracking. The above observations are summarised in Figure 5.1.

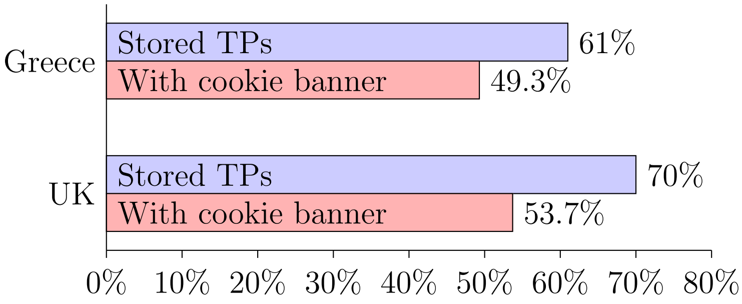


Figure 5.1: Websites that store TPs and have a cookie banner implementation.

5.2.2 Privacy Options

After the cookie banners were normalised, their privacy options were extracted based on their call to action and were manually categorised in the 4 categories introduced in Table 3.2. In total, 14,758 distinct privacy options were found. From these, 3,068 were from the Greek dataset and 11,690 belonged to the UK one. On average, Greek websites offered 2 privacy options on their cookie banners while the UK websites offered fewer than 2. Table 5.4, summarises the above findings and shows the totals and averages for both countries combined.

	Greece	UK	Both
total options	3,068	11,690	14,758
total banners	1,497	6,413	7,910
average	**2.04**	**1.82**	**1.93**

Table 5.4: The average number of privacy options that cookie banners provide in Greece and the UK.

Since all the privacy options have been properly categorised, it is possible to show the most common privacy categories that websites offer in their cookie banners. For Greece, the most common category is the Affirmative with 1,417 options (46.19%), then it is the Informational with 752 options, third is the Managerial with 596 options (19.43%) and fourth the Non-Affirmative category with only 303 identified options (9.88%).

Interestingly, the UK follows an almost identical pattern with Greece on the distribution of the cookie banner options. The most dominant categories are the Affirmative and Informational with 5,365 (48.2%) and 4,405 (37.68%) options respectively. Then, the Managerial category comes third with 1,289 options (11.03%) and finally, the Non-Affirmative category with only 361 opt-out options (3.09%). Figure 5.2, depicts the similarities of the dominant and non-dominant privacy option categories discussed here.

5.2.3 Cookie Banners Without Options

While cookie banners are used by websites to inform users about tracking activity and allow them to manage their privacy settings, a number of cookie notices offer very limited privacy options and sometimes none at all. These types of cookie banners, usually inform the user that just “by using this

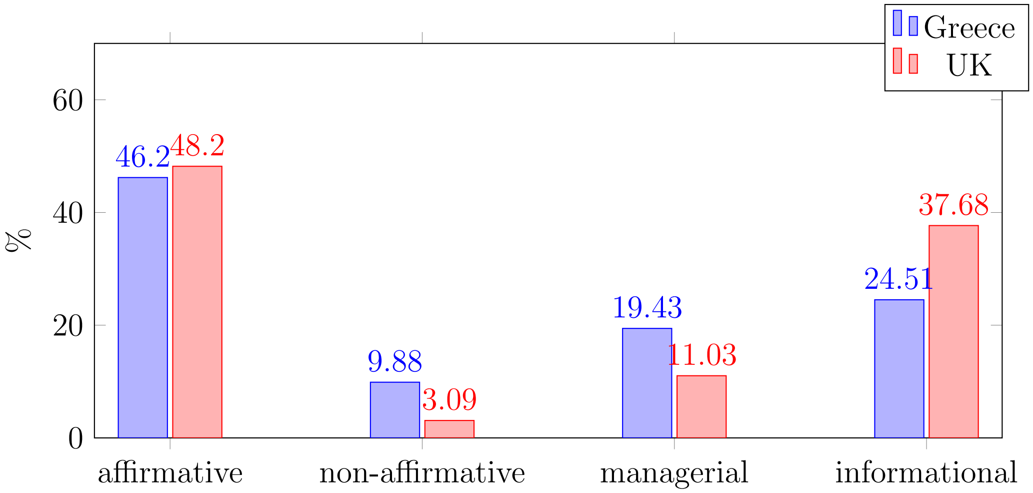


Figure 5.2: The most common privacy options offered by cookie banners in Greece and the UK.

website" they agree to be tracked and offer a dismiss button or a link to a privacy policy page or nothing at all.

However, most websites seem to offer at least one privacy option to their users. Specifically, in Greece, only 5 websites (0.33%) offer no options at all. Similarly, the UK has a low number of no-option cookie banners with only 59 websites (0.92%) that do not offer any privacy options at all.

On the other hand, single privacy options tend to be frequent among cookie banner implementations. Most commonly, cookie banners will either only offer an Affirmative option or an Informational one. More specifically, in Greece, there are 253 websites (16.9%) that offer only an affirmative option and 59 websites (3.94%) offering only a link to their privacy policy. In the UK, 1,225 websites (19.10%) allow users to only accept cookies and 563 websites (8.78%) only show a link to their privacy policy page on the cookie notice. Table 5.5, summarises the above findings.

	Greece	UK	Both
Total cookie banners	1497	6413	7910
No options	5	59	64
Only Informational	59	563	622
Only Affirmative	253	1225	1478

Table 5.5: Cookie banners that offer a single privacy option or none at all.

It is evident that the results are very similar for both Greece and the UK, as depicted by Figure 5.3. Interestingly, neither country has a website that offers a cookie banner with only a Non-Affirmative privacy option.

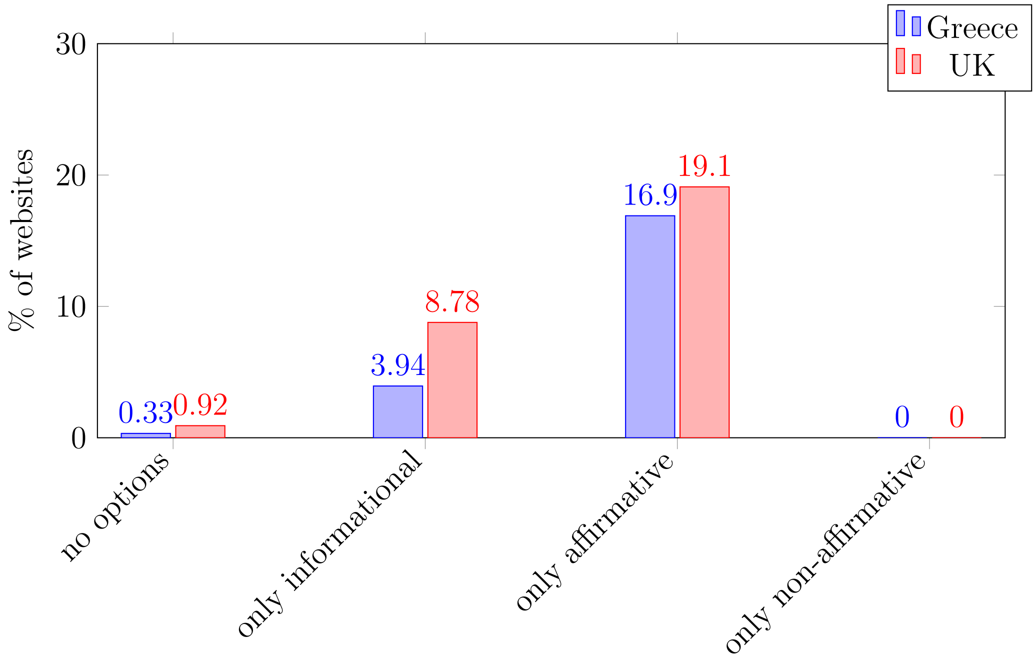


Figure 5.3: The percentage of websites offering a single option only.

5.2.4 Rejecting Cookies

An important aspect of cookie banners is to allow users to quickly and effectively reject trackers from installing cookies on their web browsers. However, it seems that a direct-opt out button has a very low adaptation among websites in both countries that were surveyed as part of this project.

More specifically, only 303 (9.88%) Greek websites offered a direct opt-out option. Similarly, only 361 (3.09%) Non-Affirmative options were identified in the UK dataset. While opt-out buttons seem to be slightly more frequent in Greece, non-affirmative privacy options are extremely rare in both countries. The above results are summarised in Table 5.6.

5.2.5 Managing Cookies

Another important aspect of cookie banners is to allow visitors to manage their privacy settings. For instance, a number of websites allow users to

	Greece	UK	Both
total options	3068	11690	14758
direct opt-outs	303	361	664
%	**9.88**	**3.09**	**4.50**

Table 5.6: The number of direct opt-out buttons offered by cookie banners.

opt-out from specific trackers.

Managerial options are significantly more prevalent compared to the Non-Affirmative ones. More specifically, Greek websites offer 596 (19.43%) cookie banners with an "options" button. Similarly, in the UK 1,289 (11.03%) Managerial privacy options were identified. Interestingly, "options" buttons appear to be slightly more frequent in Greek websites compared to UK ones. Table 5.7, summarises the above findings.

	Greece	UK	Total
cookie banners	1497	6413	7910
managerial option	596	1289	1885

Table 5.7: The cookie banners offering at least one managerial option.

5.2.6 Call to Actions

In advertising and marketing, a Call to Action (CTA) is any "device" that is designed to prompt immediate action by using an imperative verb such as "Buy Now" or "Accept All" [43]. Usually, CTAs try to create a sense of urgency to the user in order for them to act fast, so that they don't miss out on a deal [44].

Cookie banners implement buttons in order to allow users to interact with these notices and choose their privacy settings. However, the Call to Actions used by these buttons may influence a user's perception about tracking as well as their privacy choices. In order to detect such practices, this project looked at the CTAs of the cookie banners that were collected.

More specifically, a total of 1,131 unique CTAs were identified in the dataset. In Greece, there were 170 Affirmative, 58 Non-Affirmative, 83 Managerial and 170 Informational unique CTAs. Furthermore, in the UK dataset, there were 250 Affirmative, 34 Non-Affirmative, 107 Managerial and

259 Informational unique Call to Actions. Table 5.8, summarises the unique CTAs identified in the Greek and UK dataset.

	Greece	**UK**	**Total**
Affirmative	170	250	420
Non-affirmative	58	34	92
Managerial	83	107	190
Informational	170	259	429

Table 5.8: Total number of unique terms per privacy category.

Interestingly, the most common CTAs for both Greece and the UK are very similar across all 4 privacy categories. More specifically, the most common Affirmative terms in both countries are "I accept" and "Ok". Similarly, "Learn more" and "More Information" was the most common terms in the Informational category in both Greece and the UK. Figure 5.4 and Figure 5.5 depict the most common CTAs for Greece and the UK respectively.

5.2.7 Privacy Policies

In addition to privacy options, cookie banners usually contain privacy text that informs users why they are seeing the cookie notice and what cookies are used for. This text is usually very concise compared to the full Privacy Policy of the website. Since most users make privacy decisions based only on that small piece it is an important part of cookie banners and this project.

On average, the cookie banner privacy text is 59 words long in both Greece as well as the UK. More specifically, the average length of the privacy text in Greek websites is 66 words, making longer than the UK's average of 52 words. The above results are summarised in Table 5.9.

	Greece	UK	Total
cookie banners	1497	6413	7910
avg. length	66.2	52.01	59.1

Table 5.9: The average length (words) of the privacy text in the cookie banners

Although the Greek cookie banner text is longer on average, their content appears to be identical. Using the TF-IDF method, the most common terms were identified for each country. More specifically, the most common cookie banner terms were “we use” (3%), “experience” (2.96%), “better” (2.8%), “by accepting” (2.3%) and “website” (2.3%).

Interestingly, the most frequent cookie banner privacy terms in the UK are almost the same and as frequent as the Greek terms. Specifically, those are “uses” (3.5%), “best” (3%), “ensure” (2.77%), “site” (2.6%), “experience” (2.6%). The TF-IDF results and how they compare between each country are depicted in Figure 5.6.

5.3 Summary

This chapter summarised the amount of data collected before and after the crawl, including the number of websites identified for each country as well as the number of datapoints collected by OpenWPM. Furthermore, it looked at the computing resources required to run the cookie banner detection crawl and how York’s Viking supercomputer was employed to reduce runtimes.

Finally, this chapter presented in detail the results from the data analysis that was conducted on the collected cookie banners. More specifically, it demonstrated the prevalence of cookie banners, the different attributes of their privacy options as well as the TF-IDF analysis on the cookie notices’ privacy text.

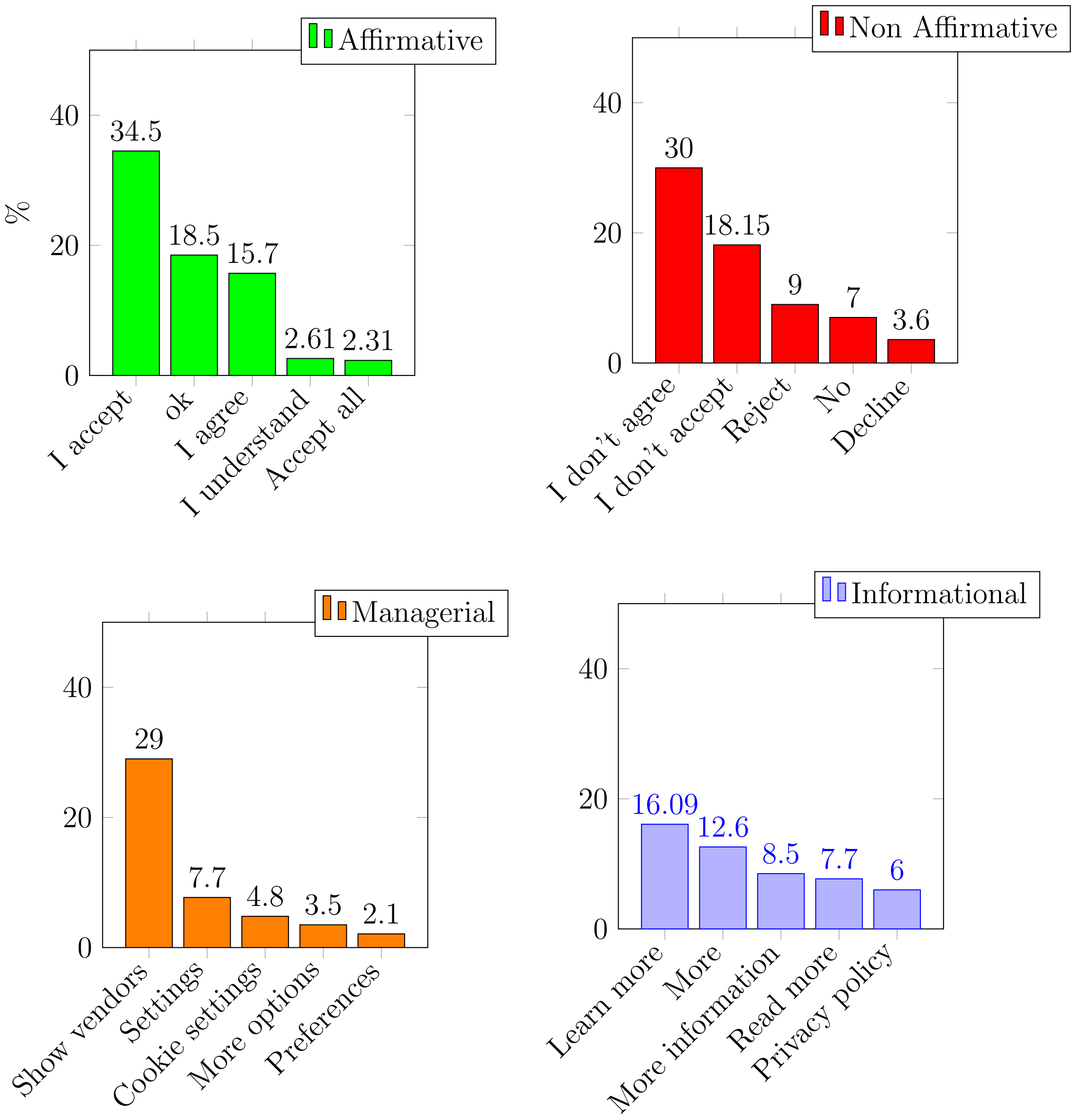


Figure 5.4: The most common Call to Actions in Greece (translated).

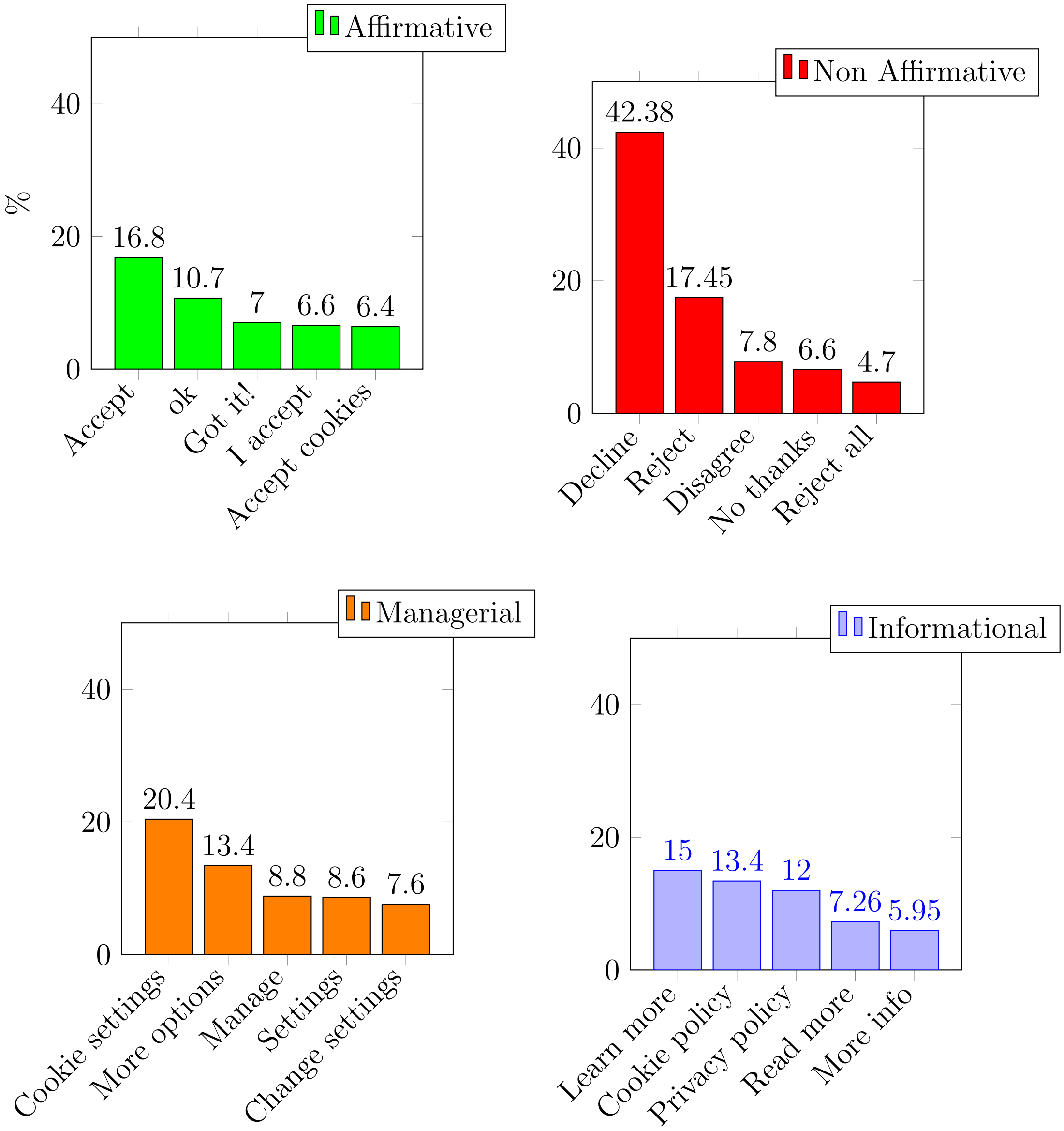


Figure 5.5: The most common Call to Actions in the UK.

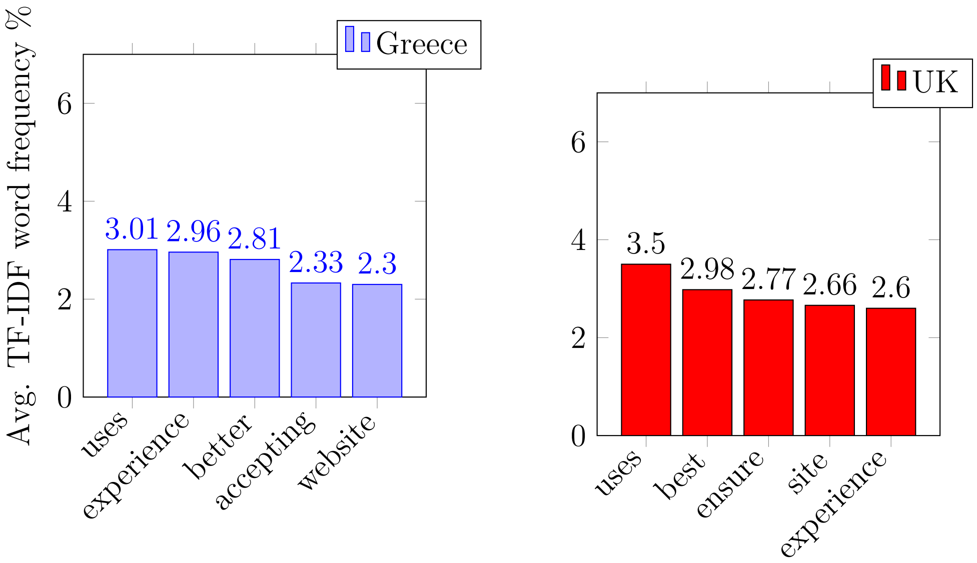


Figure 5.6: The most common cookie banner terms based on their TF-IDF values.

6 Discussion

This chapter discusses in detail the data analysis results. It will analyse their significance, the impact that cookie banner implementations have on internet users and whether they answer the research questions that were raised at the beginning of this project. This chapter can be divided into 2 parts which are summarised in the following list:

1. **Identifying dark patterns**: To identify whether websites implement their cookie banners in order to steer users towards privacy-intrusive decisions;
2. **Answering the research questions**: To establish whether the results from the cookie banner data analysis answer adequately the research questions set out in the beginning of this project.

6.1 Understanding the Results

Cookie banners might not be everywhere

Nowadays, it seems that almost every website presents its users with a cookie notice informing them that the site that they have visited is using Third-Party Cookies. Indeed, this appears to be true since half of the websites in the dataset prompt the user with a cookie banner (Greece 49%, UK 53%, see Table 5.3). These results are significantly higher than the findings of Eijk et al. [16] who detected a cookie notice on 40% of the websites that they surveyed. However, their dataset was significantly smaller and used websites from the US as well.

While this project does not make comparisons between the prevalence of cookie banners before and after the GDPR, one may look at the above data and assume that a large number of websites are complying with the EU legislation. This was observed by Degeling et al. [26] who found that more websites offered privacy policy pages after the GDPR came into force.

However, this might not be entirely true. Specifically, the findings of this project show that 61% of Greek and 70% of UK websites store at least one third-party cookie on their user's browser. As already shown, only 49% of Greek and 53% of UK websites display a cookie notice and therefore, suggesting that 15% of websites in both countries have yet to comply with GDPR (Figure 5.1).

Thus, the initial assumption of an average everyday user is that cookie banners are almost everywhere. On the contrary, it appears that a lot more work is yet to be done.

Websites really want users to accept cookies

It is clear that websites in both Greece and the UK offer approximately 2 privacy options per cookie banner (Table 5.4). Furthermore, these options are more likely going to be either only an Affirmative option or a combination of Affirmative and Informational options (Figure 5.2).

Thus, websites either only allow users to accept tracking or go through additional steps to manage their privacy settings or find out more information about how cookies are used and affect their privacy. Therefore, by hiding away information behind additional clicks and making the opt-in options more prevalent, more users will agree to third-party cookie tracking because, simply, there isn't anything else that they can do.

This can be considered as "Digital Nudging" which is defined by Weinmann, Schneider and vom Brocke as "the use of user-interface design elements to guide people's behaviour in digital choice environments" [45]. For instance, the more expensive products in supermarket shelves are placed at eye level so that customers are nudged into making unplanned purchases [46]. Similarly, websites always offer Affirmative actions and require multiple clicks for users to opt-out and thus, nudging them towards opting-into tracking.

This has been observed in a large number of websites in the dataset. Therefore, it is affecting a large number of users on a daily basis indicating a dark pattern. Interestingly, this was also observed by the Norwegian Consumer Council which found that large tech companies, such as Google and Facebook, use similar nudging techniques in order to steer users towards privacy-intrusive choices.

Opting out is hard

With only a 4.5% adaptation across Greece and the UK, it is apparent that websites do not implement opt-out buttons on their cookie banners (Table 5.6). Unless they offer an "options" button on their notice, there is no way for users to reject tracking. This can be considered a highly privacy-intrusive technique and therefore, a dark pattern.

Interestingly, the above results contradict previous research on this topic. Habib et al. [12] showed that 89% of the websites in their sample offered an opt-out option which is the complete opposite to what was found by this project. However, it is worth noting that Habib et al.'s sample consisted of only 150 websites. Furthermore, they only looked at the privacy options offered by CMPs. Therefore, all the cookie banners and their options tend to offer the same privacy options, regardless of the website.

Thus, it is evident that a larger dataset with a more diverse set of cookie banner implementations shows that direct-opt out options are not prevalent. This is supported by research from Sanchez-Rola et al. [25] They found that reject privacy options can be found in approximately only 4% of their dataset which consisted of 2,000 websites within the European Union. This dataset and results seem to be matching the findings of this project as well.

CTAs are not misleading

In total, more than 1 thousand unique Call to Actions were identified across Greece and the UK (Table 5.8). Moreover, the most common privacy category was the Affirmative and second the Informative in both countries (Figures 5.4 & 5.5).

While this aligns with the second observation made in this section, the CTAs themselves do not seem to be misleading or ambiguous regardless of how common they might be. For instance, no terms in the Non-Affirmative category were trying to give the illusion of negative consequences if the user chose to opt-out e.g: "Opt-out but checkout might not work". That was consistent across all privacy categories in both Greece and the UK.

Have some cookies, they are good for you

Every cookie banner comes with a short piece of text that usually informs the users of the purpose of that notice. While not very long or extremely specific, it is extremely important since all users will have to read it.

Specifically, the average length of the cookie banner privacy text is 60 words (Table 5.9). Surprisingly, the most common terms in both countries are almost identical and have similar frequencies (Figure 5.6). More specifically, these common terms are “experience”, “better” and “ensure”. Interestingly, terms such as “privacy” or “tracking” were not found during the data analysis.

It is clear that cookie banners aim to present the cookies as an instrument that help users enjoy a better experience while using that website. Therefore, users are inclined to click “accept”, without understanding the ramifications on their privacy. Even if the cookie banner provides an opt-out option, users might still choose to opt-in to cookies worrying that they will be unable to use that website, yet that might be very far from the truth.

The frequency that this pattern appears in both Greek and UK websites that users in both countries have to deal with such choices, and sometimes misinformation, on a regular basis. Therefore, this can be identified as a dark pattern that websites employ to steer users towards security-intrusive choices.

Both countries are very similar

Although both Greece and the UK are in the European Union and therefore, have to adhere to the GDPR they vastly differ in terms of language and size in economy and population. Thus, it can be expected that the cookie banners in each country can be different as well.

However, the findings discussed in this chapter paint a completely different picture. More specifically, the results for one country are almost identical to the other. For instance, the cookie banner privacy text TF-IDF results are almost the same in both the terms as well as statistical frequency. Another, more prominent example, is the distribution of the most common privacy options categories (Table 5.4, Figure 5.2), where the Affirmative and Informational categories are the most prevalent ones.

Therefore, it is apparent that the language or the size of a country might not matter on how users experience websites and their cookie banners. This can be attributed to two things:

1. **Mimicking**: One country simply copies the cookie banner implementation of the other. For instance, the UK has a significantly larger e-commerce industry which quickly adapted to the GDPR. Subsequently,

Greek e-commerce websites visited their UK counterparts in order to see how they complied with GDPR;

2. **CMPs**: Content Management Platforms (CMPs) offer their services across Europe. Therefore, it is possible that Greek and UK websites use the same CMPs and therefore, their cookie notices will be identical. However, this assumption requires further research to be confirmed.

6.2 Answering the Research Questions

In the beginning, this project set out 7 research questions aiming to understand cookie banners and their privacy options better. The following list discusses in detail whether the collected data and their analysis have managed to answer those research questions:

- **RQ1**: Almost half of the websites in Greece (49%) display a cookie banner while 1,871 (61%) store Third-Party Cookies. Similarly, there are 6,413 websites (53.7%) that display a cookie banner, while 8,256 (70%) of them store TPs on a user's browser;

- **RQ2**: In Greece, the average number of privacy options in a cookie notice is 2.04. Similarly, UK websites offer 1.8 options per cookie banner (Table 5.4);

- **RQ3**: Unfortunately, only 303 Greek websites ($< 10\%$) allow their users to directly opt-out from third-party tracking. In the UK, 361 websites (3%) offer an opt-out option with their cookie notice (Table 3.2);

- **RQ4**: The majority of both Greek and UK websites offer privacy options to their users. Specifically, only 5 Greek websites (0.3%) offer no options at all. In the UK, 59 websites (0.9%) do not offer privacy options (Table 5.5);

- **RQ5**: The Affirmative category is the most common in both countries. Specifically, there are 1,417 (46%) and 5,635 (48%) affirmative options in Greece and the UK respectively. Second, comes the Informational category with 752 (Greece, 24.5%) and 4,405 (UK, 37.7%) privacy options;

- **RQ6**: Overall, Greek and UK websites allow users to manage their privacy options. More specifically, 596 Greek websites (19.4%) offer a “Settings” button. In the UK, 1,289 websites (11.03%) have implemented an “Options” button in their cookie banners (Table 5.4);

- **RQ7**: The average length of the cookie banner text is 66.2 words for Greece and 52 words for the UK (Table 5.9). The 5 most frequent terms in Greece are “uses”, “experience”, “better”, “accepting”, “website”. For the UK, the 5 most common terms are “uses”, “best”, “ensure”, “site”, “experience”. The full TF-IDF term results can be found in Appendix A.3.1.

It is evident that the results adequately answer the 7 research questions. However, it is very likely that the plethora of data collected as part of this survey may be hiding more dark patterns. Thus, the dataset allows for more questions to be asked and further data analysis to be conducted.

6.3 Summary

This chapter discussed in detail the results from the cookie banner data analysis. More specifically, it looked at how websites implement their cookie banners using dark patterns. For instance, these patterns can be the lack of opt-out options or telling the users that the cookies are only used to improve their experience on the website, which is misleading. Finally, the chapter explored whether the collected data and the results adequately answer the research questions asked at the beginning of this survey. While it found, that these questions have been answered, the results may be hiding more dark patterns and therefore, more questions should be asked.

7 Conclusion

7.1 Related Work

While this project has developed novel methods and gathered a plethora of data in order to understand the cookie banner landscape better, a lot more work can still be done. This includes building upon new and existing tools as well as diving deeper into the collected data and performing further analysis of it.

Repeat the study and expand it in more countries

One of the main goals of this project was to be repeatable and easily extendible. Thus, this survey can be conducted again with minimum effort and the new results can be directly compared with the ones presented in this paper.

For instance, this project showed that approximately 12% of Greek websites do not display a cookie notice even though they store Third-Party Cookies for tracking purposes. In a few months, this survey can be conducted again in order to see whether the compliance levels have increased.

Furthermore, the crawlers developed for this project can be easily changed to target a different set of countries. For example, different researchers may want to explore different languages and countries within the European Union and beyond.

NLP in cookie banners

Natural Language Processing (NLP) is the process of analysing and understanding the large amounts of natural language text such as transcripts or books. Various NLP techniques such as Morphological, Syntactic and Relational analysis [47] can be added to this project in order to improve data collection as well as result accuracy.

The following list summarises the components that can benefit from Natural Language Processing extensions:

1. **Terms of service parser**: The current version of the TOS parser searches for specific exclusionary terms within the text in order to determine whether the website refuses crawlers. Using NLP, this process can become more accurate. For instance, more terms can be efficiently searched or the context of the entire text can be understood better;
2. **Extracting privacy option terms**: Currently, detecting privacy option terms is done manually. While this can be beneficial as local nuances can be easily weeded out, the survey can slow down due to the manual work involved. Natural Language Processing techniques can be implemented to assist human reviewers with this process;
3. **Understanding the privacy text**: The cookie banner privacy text is currently analysed using the TF-IDF method, which is also an NLP method. However, a lot more NLP analysis can be conducted in order to understand the privacy text included with the cookie banners.

Website ranking

Due to software bugs and time limitations, this project failed to store the Tranco ranking of each website that was surveyed. This can provide further information on compliance and the cookie banner landscape.

For instance, by knowing the website rankings may offer further insights such as whether rank and GDPR compliance have a correlation. Furthermore, parallels can be drawn between websites that employ dark patterns to nudge users towards privacy-intrusive decisions and their ranking.

Rise of fingerprinting

Fingerprinting is the process of adding a unique identifier (fingerprint) to an object or identifying an object from its fingerprint [48]. Browser fingerprinting techniques and algorithms have been suggested as an alternative to Third-Party Cookies for user tracking online. While cookies can be blocked or deleted by modern web browsers, browser fingerprinting methods rely on the browser's and computer's characteristics, which are usually not controlled by users, to identify a user [49]. For instance, Panopticlick gathered user information from the browser's timezone, list of plugins, fonts, the HTTP connection parameters and more [50].

Since Google is aiming to completely block Third-Party Cookies from Chrome by 2022 [51], browser fingerprinting might become the preferred method of online tracking and advertising. As Englehardt et al. [1] showed in their 1-million website survey, canvas fingerprinting was detected in 5% of their sample suggesting that websites are already aware of such methods.

Therefore, this projected can be extended as follows in order to measure the impact of browser fingerprinting:

1. Using OpenWPM, measure whether online fingerprinting has increased before and after Google has blocked Third-Party Cookies;
2. Extending the methods developed for this project where needed, assess whether browser fingerprinting had an impact on the cookie notices shown by websites.

For instance, if a significant increase in browser fingerprinting is detected in the UK, do cookie notices reflect the change in which websites track users?

7.2 Final Remarks

The rise of the internet has brought with it numerous online services such as e-banking, shopping, news and many more. At the same time, companies are trying to track the user's online activities in order to improve their advertising campaigns. Thus, while users are encouraged to read their news or pay their bills online, they are (unwillingly most of the times) also provide advertisement campaigns with their personal information.

The European Union and the UK have passed legislation, such as the GDPR and the Data Protection Act 2018, that aim to give users the power to request their data and also have them deleted. Furthermore, these laws require websites to disclose to their users their tracking activities, what information they store, how long they are storing it for and who is having access to it. These disclosures usually come in the form of"cookie banners", which are windows that pop-up when a user visits a website for the first time. In addition to the tracking information, it is common for cookie banners to contain privacy options such as "Accept/Reject Tracking".

Since cookie banners are becoming part of everyday online life, it is important to understand them better. This project aimed to build upon existing research on this topic but also shed light on the aspects of cookie banners that have not been explored before. This study focused on cookie notices in two countries namely, Greece and the UK. These were chosen for two reasons. Firstly, they are both covered by data protection legislation and therefore, there was an abundance of cookie banners. Secondly, both countries vastly differ in language as well as the size of their population and economy and therefore, differences across a user's experience in the web would be noticeable.

In order to crawl and collect cookie notices from 15,000 websites across Greece and the UK, existing tools had to be extended and new methods had to be developed. More specifically, this project developed software to identify the most popular websites in the countries of interest, using lists such as Tranco. After the websites had been identified, their robots.txt and the Terms of Service were analysed in order to verify that they allow tracking. To crawl and collect cookie banners from these, OpenWPM was used. Specifically, OpenWPM was extended in order to look for cookie banners within a website by taking advantage of the "I don't care about cookies" CSS selectors list. Once a cookie notice was found, it was stored in an SQLite database for later analysis. In total, more than 7,000 cookie banners were collected and OpenWPM gathered approximately 15,000,000 datapoints including Third-Party Cookies, HTTP Requests, scripts and many more.

Before analysing this rich dataset, it first had to be sanitised and then normalised. This was a combination of manual inspection that weeded out local nuances as well as scripts that automatically categorised the privacy options found in cookie banners and converted them into a tabular format. This project found that approximately 53% of websites display a cookie banner. However, at least 65% of websites store Third-Party Cookies and thus, it is evident that 12% of websites do not comply with the privacy laws of Greece and the UK. Furthermore, it was found that websites make it really difficult for users to reject tracking with only 4.5% offering an opt-out option. Instead, they make Affirmative privacy options a lot more prevalent. Furthermore, it is evident that websites present cookies as "devices" that improve usability and browsing experience for a user and therefore, nudging them to accept tracking even though that is not true.

In conclusion, this project aimed to develop a comprehensive understanding of the cookie banner landscape and how this may affect everyday users and lawmakers alike. It is clear that although cookie banners are everywhere, there is still a large number of websites that do not comply with the law or implement dark patterns to trick, and sometimes force, users into accepting tracking. This is noticeable in the low adaptability of opt-out options, the prominence of Affirmative options and the "branding" of cookies as the "secret" to a better browsing experience. Although the GDPR and the Data Protection Act 2018 has paved the way for users to take control of their privacy, it seems that not only websites have adapted but have also found ways of conducting large-scale data collection while the powers granted to the users by legislation seem inadequate.

A Appendix

A.1 Tables

Methods

Field	Description
privacy_text	The privacy text displayed in the cookie banner.
has_accept_btn	Whether the cookie banner has an affirmative button.
cta_accept	The call to action used in the affirmative button.
has_decline_btn	Whether the cookie banner has an non-affirmative button
cta_decline	The call to action used in the non-affirmative button.
has_info_btn	Whether the cookie banner has an informational button
cta_info	The call to action used in the informational button.
has_options_btn	Whether the cookie banner has a managerial button
cta_options	The call to action used in the managerial button.

Table A.1: The SQL schema used to store the normalised cookie banners based on the 4 distinct privacy options categories.

Implementation

Greek Phrases
Αυστηρά για προσωπική χρήση
Μόνο για προσωπική χρήση

Table A.2: The Greek phrases used as exclusion terms in the TOS parser.

Greek	English
ΟΡΟΙ ΧΡΗΣΗΣ	Terms of service
Όροι Χρήσης	Terms of Service
Όροι χρήσης	TERMS OF SERVICE
ΌΡΟΙ ΧΡΉΣΗΣ	Terms of use
όροι χρήσης	Terms of Use
	TERMS OF USE

Table A.3: Greek and English terms used to identify Terms of Service links.

A.2 Code

Due to the size of the project, the source code has been moved to GitHub. The full repo can be found here `https://github.com/george-kampanos/i-like-cookies`

A.2.1: `https://github.com/george-kampanos/i-like-cookies/blob/master/code/step1a_url_parser.py`

A.2.2: `https://github.com/george-kampanos/i-like-cookies/blob/master/code/step1b_checker_robots.py`

A.2.3: `https://github.com/george-kampanos/i-like-cookies/blob/master/code/step1c_checker_tos.py`

A.2.4: `https://github.com/george-kampanos/i-like-cookies/blob/master/code/step2a_css_selectors_parser.py`

A.2.5: `https://github.com/george-kampanos/i-like-cookies/blob/master/code/step2b_openwpm_cookie_parser.py`

A.2.6: `https://github.com/george-kampanos/i-like-cookies/blob/master/code/OpenWPM/automation/Commands/browser_commands.py`

A.2.7: `https://github.com/george-kampanos/i-like-cookies/blob/master/code/OpenWPM/automation/Commands/command_executor.py`

A.2.8: `https://github.com/george-kampanos/i-like-cookies/blob/master/code/OpenWPM/automation/CommandSequence.py`

A.2.9: `https://github.com/george-kampanos/i-like-cookies/blob/master/code/OpenWPM/automation/Commands/utils/cookie_utils.py`

A.2.10: `https://github.com/george-kampanos/i-like-cookies/blob/master/code/OpenWPM/automation/Classes/CookieBanner.py`

A.2.11: `https://github.com/george-kampanos/i-like-cookies/blob/master/code/OpenWPM/automation/Commands/Types.py`

A.2.12: `https://github.com/george-kampanos/i-like-cookies/blob/master/code/step3a_parse_cookie_banners.py`

A.2.13: `https://github.com/george-kampanos/i-like-cookies/blob/master/code/rq1.sql`

A.2.14: `https://github.com/george-kampanos/i-like-cookies/blob/master/code/rq2_3.py`

A.2.15: `https://github.com/george-kampanos/i-like-cookies/blob/master/code/rq4_5.sql`

A.2.16: `https://github.com/george-kampanos/i-like-cookies/blob/master/code/rq6.sql`

A.2.17: `https://github.com/george-kampanos/i-like-cookies/blob/master/code/rq7.py`

A.3 Spreadsheets

A.3.1: **Data Analysis Spreadsheet**: `https://github.com/george-kampanos/i-like-cookies/blob/master/data/data-analysis.xlsx`

Bibliography

[1] S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 1388–1401.

[2] R. of .gr and .el Domain Names, accessed 25/08/2020. [Online]. Available: https://grweb.ics.forth.gr/public/domains/registration

[3] "Additional domains," accessed 25/08/2020. [Online]. Available: https://www.nominet.uk/uk-domains/additional-domains/

[4] ".uk Rules of Registration," accessed 25/08/2020. [Online]. Available: https://media.nominet.uk/wp-content/uploads/2018/05/22141819/dotUK-Rules-of-Registration.pdf

[5] Accessed on 31/08/2020. [Online]. Available: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=59977

[6] Accessed on 31/08/2020. [Online]. Available: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=67075

[7] A. Hern, "Fitness tracking app strava gives away location of secret us army bases," Jan. 2018, accessed on 31/08/2020. [Online]. Available: https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases

[8] F. Roesner, T. Kohno, and D. Wetherall, "Detecting and defending against third-party tracking on the web," in *Presented as part of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*. San Jose, CA: USENIX, 2012, pp. 155–168. [Online]. Available: https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/roesner

[9] E. Graham-Harrison and C. Cadwalladr, "Revealed: 50 million facebook profiles harvested for cambridge analytica in

major data breach," Mar. 2018, accessed on 31/08/2020. [Online]. Available: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

[10] "Gdpr: Official legal text," Sep. 2019, accessed on 31/08/2020. [Online]. Available: https://gdpr-info.eu/

[11] "Data protection act 2018," accessed on 31/08/2020. [Online]. Available: https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

[12] H. Habib, Y. Zou, A. Jannu, N. Sridhar, C. Swoopes, A. Acquisti, L. F. Cranor, N. Sadeh, and F. Schaub, "An empirical analysis of data deletion and opt-out choices on 150 websites," in *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, 2019.

[13] M. Nouwens, I. Liccardi, M. Veale, D. Karger, and L. Kagal, "Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence," *arXiv preprint arXiv:2001.02479*, 2020.

[14] C. Jensen and C. Potts, "Privacy policies as decision-making tools: an evaluation of online privacy notices," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 2004, pp. 471–478.

[15] F. J. Zuiderveen Borgesius, S. Kruikemeier, S. C. Boerman, and N. Helberger, "Tracking walls, take-it-or-leave-it choices, the GDPR, and the ePrivacy regulation," *Eur. Data Prot. L. Rev.*, vol. 3, p. 353, 2017.

[16] R. v. Eijk, H. Asghari, P. Winter, and A. Narayanan, "The impact of user location on cookie notices (inside and outside of the european union)," in *Workshop on Technology and Consumer Protection (ConPro'19)*, 2019.

[17] D. Kladnik, "I don't care about cookies," accessed 27/05/2020. [Online]. Available: https://www.i-dont-care-about-cookies.eu/

[18] N. Fruchter, H. Miao, S. Stevenson, and R. Balebako, "Variations in tracking in relation to geographic location," *arXiv preprint arXiv:1506.04103*, 2015.

[19] C. Matte, N. Bielova, and C. Santos, "Do Cookie Banners Respect my Choice? Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework," *arXiv preprint arXiv:1911.09964*, 2019.

[20] C. Utz, M. Degeling, S. Fahl, F. Schaub, and T. Holz, “(Un) informed Consent: Studying GDPR Consent Notices in the Field,” in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 973–990.

[21] O. Kulyk, A. Hilt, N. Gerber, and M. Volkamer, “This website uses cookies: Users’ perceptions and reactions to the cookie disclaimer,” in *European Workshop on Usable Security (EuroUSEC)*, 2018.

[22] N. C. Council, “Deceived by design, how tech companies use dark patterns to discourage us from exercising our rights to privacy,” *Norwegian Consumer Council Report*, 2018.

[23] “Adblock plus: The world’s #1 free ad blocker,” accessed 27/05/2020. [Online]. Available: https://adblockplus.org/

[24] P. Leon, B. Ur, R. Shay, Y. Wang, R. Balebako, and L. Cranor, “Why johnny can’t opt out: a usability evaluation of tools to limit online behavioral advertising,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 589–598.

[25] I. Sanchez-Rola, M. Dell’Amico, P. Kotzias, D. Balzarotti, L. Bilge, P.-A. Vervier, and I. Santos, “Can I Opt Out Yet? GDPR and the Global Illusion of Cookie Control,” in *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, 2019, pp. 340–351.

[26] M. Degeling, C. Utz, C. Lentzsch, H. Hosseini, F. Schaub, and T. Holz, “We value your privacy... now take some cookies: Measuring the GDPR’s impact on web privacy,” *arXiv preprint arXiv:1808.05096*, 2018.

[27] J. Sørensen and S. Kosta, “Before and after GDPR: The changes in third-party presence at public and private European websites,” in *The World Wide Web Conference*, 2019, pp. 1590–1600.

[28] P. Agarwal, S. Joglekar, P. Papadopoulos, N. Sastry, and N. Kourtellis, “Stop tracking me bro! differential tracking of user demographics on hyper-partisan websites,” *arXiv preprint arXiv:2002.00934*, 2020.

[29] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, “Tranco: A research-oriented top sites ranking hardened against manipulation,” in *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, ser. NDSS 2019, Feb. 2019.

[30] "Report on the Delegation of the .ελ ("el") domain representing Greece in Greek script to ICS-FORTH GR," Oct. 2015, accessed 11/06/2020. [Online]. Available: https://www.iana.org/reports/2015/greece-report-20151005.html

[31] P. Skeldon, "Online shopping surges by 129% across uk and europe and ushers in new customer expectations of etail," Apr. 2020, accessed 31/07/2020. [Online]. Available: https://internetretailing.net/covid-19/covid-19/online-shopping-surges-by-129-across-uk-and-europe-and-ushers-in-new-customer-expectatio

[32] L. Columbus, "How covid-19 is transforming e-commerce," Apr. 2020. [Online]. Available: https://www.forbes.com/sites/louiscolumbus/2020/04/28/how-covid-19-is-transforming-e-commerce/

[33] M. Koster, "A standard for robot exclusion," Jun. 1994, accessed 11/06/2020. [Online]. Available: http://www.robotstxt.org/orig.html

[34] Seomoz, "seomoz/reppy," Mar. 2020, accessed 11/06/2020. [Online]. Available: https://github.com/seomoz/reppy/

[35] L. Richardson, "Beautiful soup documentation," *Dosegljivo: https://www. crummy. com/software/BeautifulSoup/bs4/doc/.[Dostopano: 7. 7. 2018]*, 2007.

[36] "string - common string operations," accessed 16/06/2020. [Online]. Available: https://docs.python.org/3/library/string.html

[37] "Selenium automates browsers. that's it!" accessed 11/06/2020. [Online]. Available: https://www.selenium.dev/

[38] L. Wood, A. Le Hors, V. Apparao, S. Byrne, M. Champion, S. Isaacs, I. Jacobs, G. Nicol, J. Robie, R. Sutor *et al.*, "Document object model (DOM) level 1 specification," *W3C recommendation*, vol. 1, 1998.

[39] J. S. Bowman, S. L. Emerson, and M. Darnovsky, *The practical SQL handbook: using structured query language.* Addison-Wesley Longman Publishing Co., Inc., 1996.

[40] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, 1972.

[41] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242. New Jersey, USA, 2003, pp. 133–142.

[42] “It services,” accessed 16/08/2020. [Online]. Available: https://www.york.ac.uk/it-services/research-computing/viking-cluster/

[43] B. Eisenberg and J. Eisenberg, *Call to action: secret formulas to improve online results.* HarperCollins Leadership, 2006.

[44] T. Hornor, “Writing a better call to action,” Apr. 2012, accessed 18/08/2020. [Online]. Available: http://www.marketingprofs.com/articles/2012/7772/writing-a-better-call-to-action

[45] M. Weinmann, C. Schneider, and J. Vom Brocke, “Digital nudging,” *Business & Information Systems Engineering*, vol. 58, no. 6, pp. 433–436, 2016.

[46] C. Schneider, M. Weinmann, and J. Vom Brocke, “Digital nudging: guiding online user choices through interface design,” *Communications of the ACM*, vol. 61, no. 7, pp. 67–73, 2018.

[47] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, “The stanford corenlp natural language processing toolkit,” in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.

[48] N. R. Wagner, “Fingerprinting,” in *1983 IEEE Symposium on Security and Privacy.* IEEE, 1983, pp. 18–18.

[49] K. Boda, Á. M. Földes, G. G. Gulyás, and S. Imre, “User tracking on the web via cross-browser fingerprinting,” in *Nordic conference on secure it systems.* Springer, 2011, pp. 31–46.

[50] P. Eckersley, “How unique is your web browser?” in *International Symposium on Privacy Enhancing Technologies Symposium.* Springer, 2010, pp. 1–18.

[51] D. Bohn, “Google to ’phase out’ third-party cookies in chrome, but not for two years,” Jan. 2020, accessed 02/09/2020. [Online]. Available: https://www.theverge.com/2020/1/14/21064698/google-third-party-cookies-chrome-two-years-privacy-safari-firefox