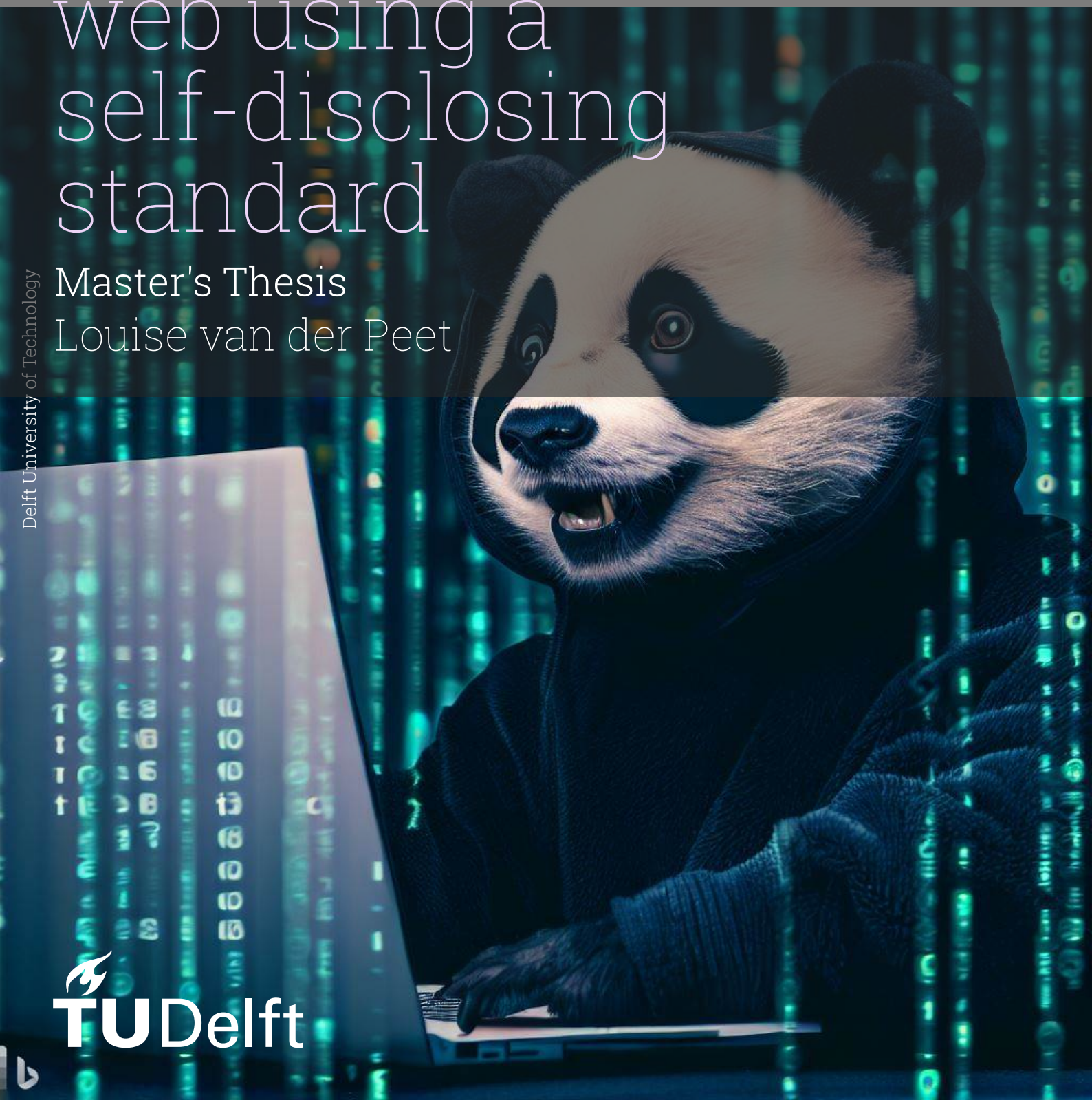
Increasing privacy-related transparency on the web using a self-disclosing standard

Master's Thesis

Louise van der Peet

Delft University of Technology



TU Delft

Increasing privacy-related transparency on the web using a self-disclosing standard

by

Louise van der Peet

Student Name	Student Number
Louise van der Peet	5641500

Instructor: G. Smaragdakis
Project Duration: December 2022 - July 2023
Faculty: Faculty of Computer Science, Delft

TU Delft

Abstract

Large amounts of trackers and other data collection forms increasingly invade users' privacy on the web. The General Data Protection Regulation (GDPR) aims to address these issues in Europe, but many violations are still made, and overall transparency is low. However, GDPR auditing frameworks and mechanisms are still missing.

We address this issue by introducing gdpr.txt: a self-disclosing privacy transparency standard. The standard uses a single reference point and machine-readable grammar to facilitate accessibility, consistency, evolvability, and, eventually, transparency of privacy-related information. Furthermore, we develop auditing tools to facilitate the automatic creation and auditing of gdpr.txt files. This includes a banner detection tool with verified accuracy of 71% and privacy policy detection with an accuracy of 80%. Then, we use these tools to gather information about the privacy landscape and find similar cookie banners, privacy policy and Consent Management Platform occurrences as in previous studies. Furthermore, we research website categories and find gambling websites have exceptionally low rates of banners and privacy policies, while news & media websites find high rates in both. We also find that cookies can differ between browsers, locations, and operating systems, making the automatic generation of cookie data difficult.

Louise van der Peet
Delft, July 2023

Contents

1 Introduction

Privacy has become a central topic in politics and daily life over the last decade. Data monetization gives companies incentive to collect information and violate the privacy of their users, causing privacy to be a larger concern than ever. As an increasing number of businesses earn their primary income using data monetization, users remain poorly informed on what happens to their data.

Due to all this, legislators have started to regulate data collection and processing on digital platforms. In Europe, the ePrivacy directive was introduced in 2009. This, among other privacy-related regulations, required data collectors to ask for consent before tracking users with cookies. However, the requirement often resulted in a pop-up or banner simply informing users about cookies without asking for explicit consent [Poullet, 2010]. In 2018, the GDPR (General Data Protection Regulation) was introduced in Europe to set a higher standard for data privacy, specifically including the requirement of stricter user consent for cookies and to include privacy policies.

However, there are still websites that do not conform to the GDPR [Dabrowski et al., 2019]. Many websites activate non-essential cookies before asking for user consent, do not have compliant banners, or do not comply to the privacy policy regulations. This makes it extremely difficult for users to know what actually happens to their data. In this thesis, we would like to propose a standard that clearly shows which cookies are used, where the privacy policy can be found, and introduce an automated audit tool to make regulation easier. This will make it far easier for users to gain agency over their personal data. The standard could also help supervisory authorities by making supervision more comprehensive, and help data protection officers to easily validate their privacy solutions. The proposed solution will be based on txt standards like ads.txt and robots.txt, which gives machine-readable information on a single references point.

1.0.1. Problem definition

Mainly due to data monetization, information privacy has become an even larger concern in the digital age. An increasing number of businesses' primary profit comes from collecting and selling their users' data. Users are often poorly informed of what happens to their data, and until recently the government did not interfere much with this breach of privacy.

To protect user's privacy better, the GDPR was introduced in Europe. However, websites often do not comply to the GDPR. This is usually due to three different kinds of reasons: interpretation or misunderstanding of the law; poor technical implementation; or economic gain.

The GDPR set out to make clear guidelines for cookie usage and privacy policies on the internet. However, the law could be interpreted differently, making it hard to implement a website that is properly compliant. As an example, the GDPR mentions that cookies should not be saved for an unnecessary amount of time. Different official sources state a different maximum time for cookie retention. For instance, an article written by the managing director of the GDPR, mentions that cookies should persist

for maximally a year [Koch, 2020], while the Dutch authority of personal data mentions that half a year is too long [Persoonsgegevens, 2017]. There are multiple examples like this, that make cookies difficult to implement for website owners who do not have expertise in law. This is why one step to protect data privacy is to make comprehensible laws and guidelines. In this research we will use the more commonly implemented one-month cut-off, which is used for distinguishing transient cookies from consistent tracking cookies [Acar et al., 2014].

Technical implementation could be an issue as well. Everyone can make a website, but not all of these people fully understand how to implement data security. Google has easy-to-use cookies, but they can be set in way that do not comply to the GDPR. Some solutions for creating ready-to-use cookie banners can be set to not comply as well, e.g. the WordPress solution Cookiebot by Usercentrics [UserCentrics, 2012] can be set to pre-tick boxes for non-essential cookies [1], which is not allowed according to the GDPR [Persoonsgegevens, 2019]. A way to solve this problem, would be to publish clear technical guidelines, with examples of proper implementation. It could also be solved by auditing vendors of cookie solutions.

Non-compliance could be caused by economic reasons. Some businesses might gain more funds from data monetization and simply paying a possible fine, than not using these types of cookies. A reason for this could be the relatively small fines. Most GDPR fines levied from 2018 to 2020 have been relatively small [Wolff and Atallah, 2021]. The fines are often also not carried out. Authorities usually do not have the budget or manpower to properly focus on imposing administrative fines [Golla, 2017].

For this research, we would like to assess the compliance of the GDPR on the web and propose a standard for transparency of user privacy. This would firstly help the interpretation of the law and technical implementation: with a standard, website owners could more easily verify whether their website is GDPR compliant. If the law may change, the standard could easily be adjusted to suit the change, and website owners could maintain the changes. The standard could help supervisory authorities to access the compliance of a website, this might also make the economic incentive smaller, as fines for wrongful data usage might become more common. Furthermore users could have more grip on their privacy because information would be more easily accessible with the implementation a transparent framework

1.0.2. Research questions

- How can we increase GDPR compliance on websites by creating a new self-disclosing standard for transparency of cookies usage and privacy policies?
  - How can we implement an automatic detection standard for cookies, privacy policies and cookie consent banners?
  - How can we implement a method to facilitate auditing of GDPR compliance on the web?
  - Is it feasible to implement a self-disclosed single reference point in website for GDPR compliance?
  - What characteristics of GDPR compliance need more attention on popular websites?

[1] Example: delta-n.nl (accessed on 15/6/2022)

2 Background

In this section we present necessary background and discuss the related work that is necessary to understand the concept of the thesis. This includes information about cookies and trackers, compliance and the GDPR, cookie banners and privacy policies, and related work. Eventually we use this knowledge to create criteria for an ideal privacy transparency framework.

2.1. Cookies and Trackers

2.1.1. Cookies

Cookies are text files that are created and stored on a user's device when they visit a website. The idea is that when the user revisits the website, this text file will be sent back to the website with the stored information. Cookies generally store user preferences and settings, and keep track of user activity on the site. This can for example be user preferences or contents of shopping carts. Cookies can even be used to store login information and other sensitive data, such as credit card numbers. [Harding et al., 2001]

Cookies can be used for a variety of purposes, including improving website functionality and performance, personalizing content and advertising, and tracking user behaviour across multiple sites. These purposes are usually divided into the following categories [Bollinger et al., 2022]:

- Necessary cookies: These cookies are essential for the website to function properly and provide basic features such as navigating between pages, accessing secure areas of the website, and enabling the website to remember user preferences and settings. Necessary cookies are typically set in response to user actions, such as logging in or filling out forms.
- Functional cookies: These cookies are used to enhance the user's experience by providing more personalized features and content. For example, functional cookies can remember the user's language preferences or the items in their shopping cart.
- Analytics cookies: These cookies are used to collect information about how users interact with the website, such as the pages they visit, the links they click, and the time spent on the website. This information is used to improve the website's performance and usability.
- Advertising cookies: These cookies are used to deliver targeted advertising to the user based on their browsing behaviour and interests. Advertising cookies are typically set by third-party advertising networks and social media platforms, and can track the user across multiple websites.

Session-based and persistent cookies

Cookies can be divided into two categories of duration: session-based and persistent. The key difference between session and persistent cookies is the length of time that they are stored in a user's browser. Session-based cookies are temporary and are deleted when the user closes their browser, while persistent cookies remain on the user's device for a longer period of time.

Persistent cookies can pose privacy concerns because they can be used to track a user's activity over an extended period of time. However, they can also be beneficial for users, as they allow websites to remember their preferences and settings without requiring them to manually set them each time they visit.

It's worth noting that both session-based and persistent cookies can be used for a variety of purposes, including authentication, personalization, and analytics. In the GDPR, persistent cookies are not specifically defined. Different official sources state a different maximum time for cookie retention. For example, `gdpr.eu` [Koch, 2020] defines the limit to one year, while the Dutch data protection authority uses the maximum of half a year [Persoonsgegevens, 2017]. In this research we consider cookies persistent when they remain on the device for 31 days or longer, as this is enough for functional purposes of cookies, while limiting the ability for long-term tracking. This one-month cut-off point has been used in previous research as well to distinguish a tracking cookie [Acar et al., 2014].

First- and Third-party cookies

Another distinction that can be made between cookies is first-party cookies and third-party cookies. First-party cookies are cookies that are set by the website that a user is visiting. These cookies primarily maintain information: user preferences, login information, and other details that are necessary for the website to function properly. For example, a first-party cookie might remember a user's language preference or shopping cart contents.

On the other hand, third-party cookies are set by domains other than the one that the user is currently visiting. These cookies are commonly used by advertisers and analytics companies to track user behaviour across multiple websites. For example, if a user visits a website that contains an advertisement from a third-party advertiser, that advertiser may set a cookie and the same advertiser could set a cookie on another website, tracking the user's history and preferences on multiple domains and instances.

The key difference between first-party and third-party cookies is the domain that sets them. First-party cookies are set by the domain that the user is visiting, whereas third-party cookies are set by a different domain. This difference has significant implications for privacy and data security, as third-party cookies can be used to track users across different websites and build a profile of their behaviour and preferences.

2.2. Trackers

Tracker cookies are a type of cookie that are used by advertisers and other third-party entities to track user activity across multiple websites. Tracker cookies are typically used by advertisers and marketers to collect data on users' browsing behaviour, such as the websites they visit, the products they view, and the searches they perform. This information is then used to create a personal profile for the user, and create targeted advertising campaigns and to deliver personalized content to users. Tracking cookies have been shown to appear in 90% of the highest traffic websites [Sanchez-Rola et al., 2019].

Tracker cookies are controversial because they can be used to collect sensitive information about users without their knowledge or consent. Some users may choose to block tracker cookies, use privacy-focused web browsers or extensions to protect their online privacy.

Session-based and persistent trackers

Tracker cookies can both be persistent and session-based. While persistent cookies are commonly used for tracking users over a longer period of time, some tracker cookies may be session cookies that are deleted when the user closes their browser.

For example, a website may use a session cookie to track a user's browsing behaviour during a single session, such as the pages they visit and the items they add to their shopping cart. This information can be used to improve the user's experience on the website by suggesting related products or services.

However, many tracking cookies are persistent cookies that are stored on the user's device for a longer period of time, sometimes up to several years. These cookies can be used to track the user's behaviour across multiple websites and to deliver targeted advertising.

First- and Third-party cookies

As mentioned before, third-party cookies can be used to track users across different websites. However, tracker cookies are not always third-party cookies. While third-party cookies are commonly used for tracking users across multiple websites, some tracker cookies may be first-party cookies.

It has been shown that 97.72% of the websites have first-party cookies that are set by third-party JavaScript [Chen et al., 2021]. The first-parties can read or set any of the third party code, making data leakage and tracking through these first-party cookies possible as well.

Many websites contain tracker and persistent cookies. Tracker cookies can even be hidden behind third party cookies or other hiding practises [Fouad et al., 2018]. This indicates a concerning privacy landscape and need for transparency.

2.3. Compliance and the GDPR

2.3.1. GDPR and Cookie Consent

The GDPR is a data protection law that became effective in May 2018 in the European Union. It grants individuals more control over their personal data and seeks to harmonize data protection regulations across EU member states. The GDPR applies to organizations processing personal data of EU residents, irrespective of their location. It introduces principles such as transparency, purpose limitation, and lawfulness for data handling. Non-compliance with the GDPR can result in significant fines, prompting organizations to prioritize responsible and secure data management practices. The GDPR has several articles and recitals related to cookie consent and privacy policies. In this section we discuss which are the most important regarding cookie consent.

Article 4(11) and Article 7: Definition of consent

Article 4 of the GDPR establishing common understanding for concepts used throughout the regulation. One of which is consent, which article 4 defines as following:

> "'Consent' of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her."
> [Consulting, 2020b]

According to article 7 [Consulting,] and Recital 32 [Consulting, 2020d] of the GDPR, this means that in order for consent to be valid, it must be:

- Freely given: The data subject must have a real choice and not be forced or coerced into giving consent.
- Specific: The consent must relate to a specific purpose, and should not be vague or general.
- Informed: The data subject must be informed of the identity of the controller, the purposes of the processing, the types of personal data being processed, and other information necessary to make an informed decision.
- Unambiguous: The consent must be given through a clear and affirmative action, such as ticking a box or clicking a button, and must not be implied or assumed.

Recital 32 also emphasizes that users should have the ability to easily withdraw their consent for cookies. It states that it should be as easy to withdraw consent as it is to give it. Users should be informed about their right to withdraw consent and provided with clear instructions on how to do so.

Article 6(1)(a): Lawfulness of processing - Consent

Article 6(1)(a) of the GDPR sets out one of the six lawful bases for processing personal data, which is “the data subject has given consent to the processing of his or her personal data for one or more specific purposes.“ [Consulting, 2020c]. This means that processing personal data is allowed under the GDPR if the data subject has given their consent to the processing, and the processing is for one or more specific purposes that have been communicated to the data subject.

It’s important to note that consent is not the only lawful basis for processing personal data under the GDPR, article 6. The other lawful bases in for processing are:

- Contractual necessity: Processing is necessary for the performance of a contract with the data subject.
- Legal obligation: Processing is necessary for compliance with a legal obligation.
- Vital interests: Processing is necessary to protect the vital interests of the data subject or another person.
- Public interest: Processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority.
- Legitimate interests: Processing is necessary for the legitimate interests of the controller or a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject.

Controllers must carefully consider which lawful basis for processing applies in each case, and ensure that they have a valid legal basis for processing personal data in compliance with the GDPR.

Article 13: Purposes of processing

In the context of cookie consent, Article 13 [Consulting, 2020a] of the GDPR requires organizations to provide individuals with specific information regarding the purposes of processing their personal data, including the purposes of any cookies or similar technologies used on their website.

This means that when obtaining consent for the use of cookies, organizations need to clearly communicate to individuals the reasons why their personal data is being processed through the use of cookies. This information should be provided in a transparent and easily understandable manner, ensuring that individuals are informed about the specific purposes for which their personal data will be processed when they accept the use of cookies.

In practical terms, this requirement implies that organizations should include clear explanations in their privacy policies or cookie banners, detailing the purposes for which cookies are used on their website. It helps individuals make informed decisions about whether they want to accept the use of cookies or not, based on a clear understanding of the processing activities associated with them.

Cookie Usage after the GDPR

Multiple studies have charactarised cookie usage after the GDPR. On EU websites, on average the number of third parties dropped by more than 10% after the installment of the GDPR [Hu and Sastry, 2019]. The study also finds that non-EU websites have less cookie notices on average.

Furthermore results by Sanchez-Rola et al. [Sanchez-Rola et al., 2019] show that the US is similarly affected by the GDPR as Europe. This is most likely because if US websites want to offer their services to EU citizens, they still have to comply to the GDPR. Cookie notices appear in 32% of US websites against 57% of EU websites according to the same research by Sanchez et al.

2.4. Stakeholders

The stakeholders of the GDPR have been defined in “Towards an Understanding of Stakeholders and Dependencies in the EU GDPR“ by Huth et al [Huth et al., 2018]. Here, five stakeholders are defined:

the data subjects, the controller, the processor, the data protection officer and the national supervisory authority. All of these groups face different consequences and problems when it comes to data privacy and the GDPR. We discuss their roles and how this research could affect them.

The data subject is the subject whose personal data is collected. We might refer to this entity as the user, and their data as user data. This person has their privacy at stake when it comes to the subject of data management and the GDPR. The data subject might be the most important stakeholder: as privacy is a human right according to the UN [Assembly et al., 1948], and relates closely to their overall freedom. Some experts even suggest that extensive collection of personal data might lead to Orwellian societies, where citizens are being constantly surveyed [Schneier, 2015]. This could be prevented by proper privacy laws, and overall data protection. The data subject will also receive more transparency about what is happening to their personal data. The implementation of this thesis could give more direct transparency about user data to the data subject, while creating an improved landscape for their data privacy and protection.

The controller is the entity that is accountable for lawful data processing, and determines the purpose and means of data processing. In our case, this is often the website owner. This entity, either a person or a company, maintains the website and therefore the cookies, privacy policy, and overall data collection. They will also be accountable by the GDPR for non-compliance. The controller is a stakeholder in this specific research, as they could benefit from a standard and comprehensible guidelines to let them easily and properly implement data processing and privacy regulations on their websites, without risking fines

The processor processes the data of the data subject, but has no direct contact with them. In the case of cookies, this is usually a third party which buys the collected data. The stake that this entity has in our research is simply that more compliance to the GDPR could limit or change the user data that they can process.

The data protection officer is an entity without a processing or controlling unit, who is in charge of data protection, or in this case GDPR compliance. This is usually a compliance team within a company, it could also be a consultant. Furthermore, the data protection officer will serve as a contact point for the national supervisory authority. This entity could benefit from this research, as we intend to make it easier to implement a GDPR compliant website, and the data protection officer could use our standard to verify compliance. This entity could be in charge of implementing the standard.

The national supervisory authority is in charge of creating the laws and guidelines for data protection, as well as monitoring and enforcing the application. In the Netherlands, this authority is the Autoriteit Persoonsgegevens (AP). We will also refer to this stakeholder as auditor. Their stake lies in increased transparency, but even more in simplified auditing of compliance in websites.

2.5. Cookie Banners and Privacy Policies

Consent and other key concepts of the GDPR are usually implemented on websites through two popular means: Cookie Banners and Privacy Policies. In this section we will explain both cookie banners and privacy policies, and how they related to the GDPR.

2.5.1. Cookie Banners

Cookie banners are a common method used by websites to comply with the GDPR's requirements regarding cookie consent. When users visit a website, a cookie banner typically appears, informing them about the use of cookies and seeking their consent.

Cookie banners should provide clear and transparent information about the types of cookies used, their purposes, and any third parties involved in processing the data. Users should have the ability to accept or reject cookies based on this information. This is according to the GDPR principles that

consent should be freely given, unambiguous, specific and informed.

However, when websites implement cookie banners they often do not comply to the standard of unambiguous consent. Researchers have shown that cookie banners often use “dark patterns“, even after the GDPR [Hausner and Gertz, 2021] [Krisam et al., 2021] [Nouwens et al., 2020]. Dark patterns are used to guide users into favorable behaviour for another stakeholder. In the case of cookie banners this can be seen all around the web. Data processors try to guide users to accept as many cookies as possible. For example, according to Krisam et al. around 80% of cookie banners require more clicks for rejecting cookies than accepting cookies, and around 75% uses some form of visual nudging towards acceptance.

The GDPR discourages the use of dark patterns because they undermine the principles of transparency and fairness. Consent obtained through manipulative or deceptive techniques is considered invalid under the GDPR. The regulation requires that consent be freely given, without any form of coercion or deception.

2.5.2. Privacy Policies

Privacy policies are legal documents or statements that inform individuals about how their personal data is collected, used, and processed by an organization. Privacy policies are essential for transparency and compliance with data protection laws, including the GDPR. Even though the GDPR does not literally use the term privacy policy, the following regulation from article 12 emphasises the need for a privacy policy:

> “The controller shall take appropriate measures to provide any information (...) relating to processing to the data subject in a concise, transparent, intelligible and easily accessible form, using clear and plain language, in particular for any information addressed specifically to a child. The information shall be provided in writing, or by other means, including, where appropriate, by electronic means.“
> [GDPR.eu, 2020]

Under the GDPR, privacy policies should provide detailed information about the data controller’s identity, the purposes and legal bases for processing personal data, data retention periods, data subject rights, and any third-party sharing of data. Additionally, privacy policies should explain how individuals can exercise their rights under the GDPR, such as the right to access, rectify, and erase their personal data. This is mostly part of article 13 of the GDPR [Consulting, 2020a].

After the GDPR, the number of privacy policies on the top 500 websites in 2019 increased to 85% which is 16% more than a year before in 2018 [Degeling et al., 2018]. Other studies have found that 70 to 80% of US websites contain privacy policies [Liu and Arnett, 2002] [Nokhbeh Zaeem and Barber, 2017].

2.6. Consent Management Platforms and Consent in the Wild

Consent Management Platforms (CMPs) are digital tools that help websites and apps comply with data privacy regulations. They provide a standardized and user-friendly approach to obtaining and managing user consent, which enables website operators and app developers to provide transparency and control to users over their personal data. This is often applied by using a cookie banner, and often a piece of software which automatically detects cookies on websites. CMPs have become a popular and sometimes even essential for website operators and app developers who want to maintain user trust and demonstrate compliance with privacy regulations like the GDPR.

According to a study by Degeling et al. [Degeling et al., 2018], approximately 60% of European websites have some form of consent notice. Bollinger et al. [Bollinger et al., 2022] found that only 3.5% of websites use CMPs, when looking at the Alexa’s top 1 million websites. Similarly, Sanchez-Rola et al. [Sanchez-Rola et al., 2019] show that only 5% of US websites deploys CMPs and 3% of websites in

the EU. They also state that CMPs are practically not used in other parts of the world.

Research conducted by Kampanos et al. [Kampanos and Shahandashti, 2021] revealed that only 44% of approximately 14,000 websites in the UK and 48% of approximately 3,000 websites in Greece displayed a cookie banner to the user. Given that roughly 90% of all websites employ tracking cookies [Solomos et al., 2019], this indicates that many websites are failing to comply with the GDPR. Furthermore, even websites that utilize CMPs often do not fulfill their promises and violate basic rules. Nouwens et al. [Nouwens et al., 2022] found that out of 680 examined websites using a CMP, 88.2% failed to meet at least one of three simple requirements (reject as easy as accept, no pre-checked boxes, and no implied consent). Matte et al. [Matte et al., 2020] discovered that, among 1,426 selected websites, 9.89% recorded affirmative consent before users made a choice, 2.66% did not allow any cookies to be rejected, and 1.89% registered positive consent even when users rejected it. Previous studies have also shown that many CMPs attempt to influence visitors into accepting all cookies, employing tactics like nudging, which is considered a dark pattern. For instance, Utz et al. [Utz et al., 2019] observed that 57.4% of 1,000 examined websites used nudging, which involved emphasizing the “Accept All“ button or concealing the option to reject consent.

Unfortunately, the situation does not seem to be improving, as highlighted by Kampanos et al. [Kampanos and Shahandashti, 2021] While high-profile violations may face penalties, the enforcement of GDPR regulations regarding cookies lags behind, as seen in many recent studies, including the ones mentioned here.

2.7. Related work

In this section we aim to discuss similar solutions to cookie transparency on the web, as well as look into ads.txt, a different kind of transparency framework. Looking into the pro’s and cons of these frameworks can aid in creating criteria for transparency frameworks overall.

2.7.1. Ads.txt

Ads.txt [iab tech lab, 2017] is a text file designed by the Interactive Advertising Bureau (IAB) to reduce programmatic ad fraud, and increase transparency. This fraud is caused by the design RTB protocols at ad exchange auctions, which lack ways to guarantee the identity of publishers. This could, for example, result in publishers selling advertising space on a website where they are not authorized to sell. With `ads.txt` bidders could cross-verify bid requests.

Ads.txt can be implemented and self-disclosed by website owners. It is initiated by including an ads.txt text file at the root of the website. This makes the file easily accessible, by navigating to `www.websitename/ads.txt`. The file contains a list of the authorized sellers, their identification code, and the publisher’s domain name.

The ads.txt standard promotes transparency in the digital advertising ecosystem. It allows publishers to disclose which entities are authorized to sell their inventory, ensuring that advertisers and buyers can verify the legitimacy of the sellers they are dealing with and prevent fraud.

The standard has gained significant adoption within the advertising industry. Many programmatic platforms, demand-side platforms (DSPs), and exchanges have implemented support for ads.txt, making it easier for publishers to adopt and benefit from its advantages.

Ads.txt has previously been empirically studied by Bashir et al [Bashir et al., 2019], where they performed a longitudinal analysis of the standard. They found that 60% of the Alexa Top-100K websites implemented the standard.

Other .txt projects

Ads.txt is not the only standard of this nature. There are three more: security.txt, robots.txt and humans.txt with the following [Anderson, 2023]:

- security.txt: describes the process for security researchers to report vulnerabilities. It includes: contact information, a disclosure policy, and acknowledgement of researchers who report vulnerabilities.
- robots.txt: provides instructions to web crawlers. With this file a website owner can restrict web crawler's access to certain URL's.
- humans.txt: intends to give credit to the people behind a website. It includes: attribution of people, technologies and licensing.

These standards have also been implemented in different extends. Robots.txt has been implemented in around 45% of the Fortune top 1000 websites in the United States, and around 35% in the European Union [Sun et al., 2007]. However, only 0.5% of the top million websites adopt security.txt [Findlay and Abdou,], and there are no similar statistics for humans.txt currently available.

2.7.2. Do Not Track

The Do Not Track (DNT) standard is an effort to enhance user privacy on the web by allowing individuals to express their preference to opt out of online tracking and targeted advertising. It aimed to provide users with a mechanism to communicate their desire for privacy to websites and online services, with a simple and standardised method. [Kellett, 2021]

The DNT standard was implemented through the use of HTTP headers. When enabled by browser and user, the browser would send a DNT header with each request, indicating the user's preference not to be tracked. Websites and services were expected to respect this header and refrain from collecting or using user data for targeted purposes.

However, despite its initial promise, the DNT standard is no longer widely used or effective. There are several reasons for this:

- Lack of enforcement and compliance: One of the main reasons for the decline of the DNT standard is the lack of enforcement mechanisms and widespread compliance. The standard was voluntary, meaning websites and services were not legally required to honor the DNT signal. There were no repercussions like fines set by authorities. As a result, many companies chose not to implement support for DNT or ignored the signal altogether. [Hill, 2018]
- Ambiguity and interpretation: The DNT standard faced challenges in terms of interpretation and ambiguity. The specification lacked clear guidelines on how websites and services should respond to the DNT signal. The stakeholders of DNT could never come to an agreement of what a website should actually do after receiving a DNT request [Hill, 2018]. This led to inconsistent and varied interpretations, making it difficult for users to trust that their preference would be respected, and even giving them a false sense of privacy.
- Limited industry support: Despite early interest and support from privacy advocates, the DNT standard failed to gain sufficient support from key stakeholders within the advertising and online tracking ecosystem. Many advertisers, data brokers, and online platforms were reluctant to adopt DNT due to concerns about its potential impact on their business models and revenue streams. The standard has been said to "kill online growth" [Wheeler, 2012].

The Do Not Track (DNT) standard aimed to empower users with greater control over their online privacy. However, its lack of enforcement, limited industry support and ambiguous implementation have led to its decline and reduced effectiveness in the digital landscape.

2.7.3. Platform for Privacy Preferences

The Platform for Privacy Preferences (P3P) [W3C, 2001] standard is developed to enhance user privacy by providing a standardized mechanism for websites to communicate their privacy practices to users in a machine-readable format.

The P3P standard is designed to address the complexity and opacity of privacy policies on the web. It aimed to enable users to make informed decisions about sharing their personal information by

providing a standardized format for websites to express their privacy practices. P3P used a machine-readable format based on XML to represent privacy policies. Websites could publish their policies in this format, allowing user agents (such as web browsers) to automatically interpret and analyze them. Users could configure their browsers to match their privacy preferences with the policies of visited websites [W3C, 2002]. In 2007, P3P was implemented in 10% of the sites returned in the top-20 results of typical searches [Cranor et al., 2008].

P3P aimed to simplify the process of understanding and comparing privacy policies, but it is no longer widely used due to several reasons:

- Complexity and cost: One of the main reasons for the decline of the P3P standard was its limited adoption by websites and user agents. Many websites did not implement P3P due to the perceived complexity and cost associated with creating and maintaining machine-readable privacy policies. Additionally, developers are prone to make errors when implementing the standard, in 2007, around 70% of the top websites had errors in their P3P policies [Cranor et al., 2008].
- Accuracy and trustworthiness: P3P relied on self-reported information provided by websites. This led to concerns about the accuracy and trustworthiness of the privacy practices stated in P3P policies. Critics argued that the standard did not effectively address the challenge of verifying whether websites actually adhered to the policies they published. [Reidenberg and Cranor, 2002]
- Shift in privacy landscape: Rather than relying solely on machine-readable privacy policies, the focus has shifted towards obtaining explicit user consent for data collection and processing activities. Modern privacy frameworks emphasize transparency, user choice, and consent mechanisms, which go beyond the scope of what P3P was designed to address. [Grimm and Rossnagel, 2000]

The P3P standard aimed to simplify privacy policy understanding and comparison. However, its limited adoption, concerns about accuracy and trustworthiness and the evolving privacy landscape have contributed to its decline and decreased relevance in contemporary privacy practices.

2.7.4. IAB consent framework

The IAB (Interactive Advertising Bureau) consent framework [Europe, 2021] is a widely adopted industry standard that aims to provide a mechanism for obtaining and managing user consent for online advertising and data processing activities. It offers guidelines and technical specifications for publishers and advertisers to ensure compliance with the GDPR, and help the digital advertising industry to: “interpret and comply with EU rules on data protection and privacy – notably the GDPR“ [InteractiveAdvertisingBureau, 2018]

The framework provides a standardized approach to gather and transmit user consent preferences regarding the use of cookies, data collection, and targeted advertising. It aims to establish a common language and technical infrastructure for stakeholders in the digital advertising ecosystem. The framework also involves the use of CMPs, which facilitate the collection, storage, and transmission of consent signals between publishers, advertisers, and technology vendors.

Despite its adoption and benefits, the IAB consent framework is not without its challenges. Here are a few potential disadvantages:

- User Experience: Some critics argue that the consent framework can result in a poor user experience. The consent banners or pop-ups can be seen as intrusive, and the granular consent options may overwhelm users with too many choices, leading to consent fatigue or apathy. In a research on IAB consent framework banners, the following was found [Matte et al., 2020]:
  - No choice to refuse: In some cases, positive consent is stored before user choice or there is no option to refuse consent at all. This violates the freely-given consent value of the GDPR.
  - Pre-selected choices: the banner selects some cookie categories or vendors by default, where the boxes are already ticked, or sliders set to accept. This violates unambiguous consent.
  - No respect of choice: positive consent is sometimes stored even when the user refused to give consent. This violates free given consent as well.

- Dependence on CMPs: The framework relies on Consent Management Providers to facilitate consent collection and transmission. This dependence on third-party providers raises concerns about data security, transparency, and potential conflicts of interest between CMPs and other stakeholders.
- Complexity: The framework can be complex to implement due to the technical specifications and integrations required, especially for smaller publishers or advertisers with limited resources. Some implementation strings have unclear semantics, which makes it harder for third party developers to implement or rely on these. [Matte et al., 2020]

2.7.5. CookieBlock

The CookieBlock method was created in 2022 in the paper Automating Cookie Consent and GDPR Violation Detection [Bollinger et al., 2022]. The tool aims to enforce user consent without the consent banner, by using machine learning. The consent is enforced on the user side by classifying each cookie in their category and assessing this against the user-defined cookie policy. When the cookie does not match the user's cookie policy, it will automatically be deleted. The user can also make exceptions for specific domains or create a custom cookie category.

The CookieBlock program is easy to install for end-users, by installing a browser extension that then will be used for any website that the user visits in that browser. Futhermore, the solution does not need any support from industry or companies and can be fully implemented by the user to take control of their own privacy.

The paper names some limitations to the method, and some other properties might make the CookieBlock solution less suitable than other solutions:

- Loss of website functionality: because cookies - essential or not - might sometimes effect website functionality. In the study that was done 15% of websites broke in some way: either by reappearing cookie banners or login issues.
- Inability to prevent cookie creation: The approach cannot prevent cookie creation within the WebExtension API and can only remove cookies after they have been stored in the browser. This is a disadvantage that comes with implementing fully on the user side: cookies have to be deleted but will not be prevented.

2.7.6. Criteria for web privacy transparency frameworks

We defined a web privacy transparency framework as the following, and consider DNT, P3P, and the IAB Consent Framework among these:

> A web privacy transparency standard refers to a set of guidelines, practices, or protocols that aim to enhance transparency and user control over their personal information when interacting with websites and online services. These standards are designed to provide organisations and individuals with clear and accessible information about how their data is collected, used, and shared, empowering them to make informed decisions about privacy, and eventually creating an online environment that is transparent about data collection and user privacy.

In order to create an improved privacy transparency framework, we first set up criteria that such a framework should have. These criteria are based on the related work, focusing on the benefits and disadvantages of previous frameworks. We specify seven criteria:

- **Accessibility**: This criterion emphasizes making privacy information easily accessible to users. It involves ensuring that any documented information is readily available, preferably in a standardizes way. Accessible privacy information make it easier for users to understand how their data is being collected, used, and protected, and for supervisory authorities to perform auditing.
- **Machine-readability and Consistency**: Machine-readability refers to the ability of computers and software to automatically process and interpret privacy-related information. This criterion

involves structuring privacy-related information in a format that can be easily understood by machines, enabling automated analysis and comparison of privacy practices across different websites and services. Machine-readable standards facilitate the development of privacy-enhancing tools and technologies. The machine readable format goes hand-in-hand with consistency, which focuses on a uniform approach to privacy transparency across different websites and services. The result of these criteria is standardized terminology, formats, and practices to ensure that privacy information is presented in a consistent manner. Consistent privacy standards and machine readability enable users and auditors to understand and compare privacy practices more easily, fostering trust and informed decision-making.

- **Accountability**: Accountability emphasizes the need for organizations to take responsibility for their privacy practices. It involves providing clear information about who is collecting and controlling user data, as well as mechanisms for users to exercise their privacy rights and seek recourse for any violations. Transparent accountability mechanisms build trust and allow stakeholers to hold organizations accountable for their data handling practices.
- **Evolvability**: Evolvability focuses on the adaptability and flexibility of privacy standards over time. As technology and privacy concerns, as well as privacy laws evolve, it is essential for privacy transparency standards to be able to accommodate changes and advancements. Evolvable standards provide a framework for ongoing improvements and adjustments to privacy practices, ensuring that users' privacy needs are adequately addressed.
- **Industry support**: Industry support refers to the participation and adoption of privacy transparency standards by the businesses that are stakeholders. It is crucial to have widespread acceptance and implementation of these standards to ensure their effectiveness and impact. Strong industry support helps create a consistent and reliable ecosystem for privacy transparency, benefiting both users and organizations.
- **Accuracy**: Accuracy entails providing precise and up-to-date information about privacy attributes. Organizations should strive to ensure that their privacy policies and disclosures reflect their actual data collection, usage, and protection practices. Transparent and accurate information builds trust with users and helps them make informed decisions regarding their privacy. It is important that a privacy standard reflects this, and incentivises organisations to be accurate.

These criteria collectively contribute to the development of a robust web privacy transparency standard that empowers users with clear and accessible information about their data privacy and facilitates better understanding and control over their personal information, while creating a strong base for auditing privacy on the web.

2.7.7. Gdpr.txt advantages

The comparison of the gdpr.txt method, which will be explained in-depth in Section 3.2, can be found in table 2.1. We classify the complexity as low, as the implementation of the protocol simply involves creating, which can be manual but also automated, as simple text file and hosting this on the website. As the framework is based on ads.txt and robots.txt, which give high accountability to the industry, we believe that gdpr.txt could become a framework that promotes accountability on the web. However, as there is currently no industry support and the framework itself does not hold organisations accountable, the accountability is currently low.

Method	Implementation side	Type of data	Machine Readable	Complexity	Accountability	Goal
Ads.txt [iab tech lab, 2017]	Server	Ad data	Yes	Low	High	Transparency
DNT [Soghoian, 2011]	User & Server	Tracking cookies	N.A.	High	Low	Privacy
P3P [W3C, 2001]	Server	Privacy policies	Yes	High	Low	Transparency
IAB consent [Europe, 2021]	Server	Privacy Policies	No	High	Low	GDPR Compliance
CookieBlock [Bollinger et al., 2022]	User	Cookies	N.A.	Low	Low	End-user privacy
Gdpr.txt	Server	Cookies, banners, privacy policies	Yes	Low	Low	Transparency

Table 2.1: gdpr.txt compared to other standards

Overall a large advantage of gdpr.txt is that it is created with transparency as the main goal. Standards are often adopted as transparency frameworks when they are not originally meant like that. For example, the IAB consent framework is made for compliance, more than transparency. This causes the framework to be less user-friendly and mainly focus on the lawful requirements for consent and ease of implementation by organisations. Furthermore, the gdpr.txt framework requires no implementation on the user side. This increases user-friendliness and puts the responsiblity of privacy on the organisation's side. Additionally the framework does not interfer with browser functionality, as it does not change any actual cookie settings but rather focuses on the transparency. The standard also increases transparency on the most important facets of GDPR-related compliance: the cookies, banner, and privacy policy.

The advantages of gdpr.txt can be summarized as following:

1. Machine Readable: the standard can be parsed and read automatically, making automation, auditing and generation simple.
2. Easy to adopt: the simple grammar, format, and ability of auto-generation, make the standard easy to implement and adopt.
3. No change on user side: as the user does not have to interfere, privacy and transparency will be enhanced in a user-friendly manner.
4. Does not interfere with browser functionality: because gdpr.txt is only a file that is hosted on the website, it does not interfere with browser functionality.
5. Complete in most important parts of GDPR compliance: gdpr.txt aims to provide complete information about cookies, banners, and privacy policies. This is most important for GDPR compliance on the web.
6. Potential for high accountability: as other .txt standards have shown, these types of standards can lead to high accountability of stakeholders.
7. Transparency as priority: gdpr.txt is made with transparency in mind. This transparency-first method could make it more capable of providing privacy transparency.

3 Methodology

In this chapter we aim to describe the methods used to create and analyze the gdpr.txt standard. We first discuss the high level structure of the methods, then we will discuss the gdpr.txt grammar in-depth, and furthermore we will discuss the software that is used to create and verify gdpr.txt files. The code can be found on Github [1].

3.1. Requirements

We aim to make a framework that increases overall privacy transparency on the web, can be used for auditing, and also automated generation for website owners. In order to achieve this, we standardize reporting of cookies, banners, and privacy policies in a single-reference point text file.

The gdpr.txt grammar consists of standardized fields to define cookies, banners and privacy policies of a website. This file can be parsed into a database, and then easily compared to other databases of real-time cookie collections, or gdpr.txt files. We developed a set of tool for the creation and auditing of gdpr.txt files. The structure of the tools can be seen in figure 3.1.

3.2. Gdpr.txt grammar

The gdpr.txt file intends to make an easy machine-readable and -creatable format for more transparency for cookies and privacy policies. Therefore the gdpr.txt grammar should be easily machine-readable, and include the most important information about cookies and the privacy policy for both the user and the auditor.

The grammar is based of off ads.txt, which uses each line as a separate record. Gpdr.txt uses lines for separate records as well. The lines can have three types of formats, namely:

$$< FIELD\#1 >, < FIELD\#2 >, < FIELD\#3 >, < FIELD\#4 >, < FIELD\#5 >, < FIELD\#6 > < FIELD\#7 >$$

or

$$< FIELD\#1 >, < FIELD\#2 >$$

or

$$< FIELD\#1 >$$

The first of which indicates a cookie records, second a banner records, and third a privacy policy record. The cookie records contain seven predefined fields, the description per field can be found in table A.1. The banner records contain the URL of the visited website, and a boolean variable on whether there is a banner with consent options. The privacy policy record is a link to the web-page where the privacy policy is located. An example of a gdpr.txt file can be found in figure 3.2.

[1] https://github.com/kokosnoob/gdpr.txt-tools/tree/master

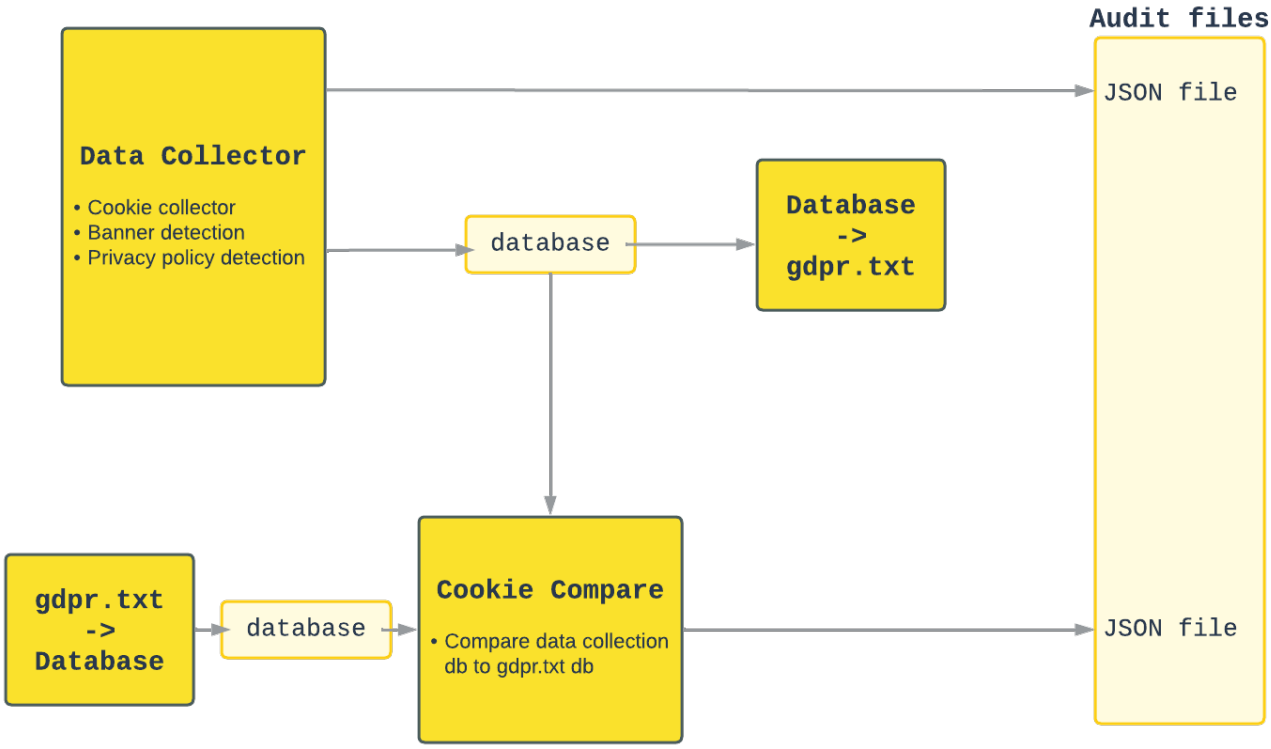


Figure 3.1: Structure of gdpr.txt audit and generation tools implementation

The grammar is designed to be machine-readable, in order to be easily scraped from the web for auditing of the GDPR. Furthermore, it is easy to implement: the features are non-ambiguous and all the data in the gdpr.txt file is already available on the website through the cookie storage.

The full implementation guide of gdpr.txt can be found in Appendix A.

3.2.1. Privacy transparency framework criteria

In Section 2.7.6 we discuss the criteria that a privacy transparency framework should have. In this section we will evaluate the gdpr.txt standard against the established criteria.

Accessibility

Privacy transparency should be for everyone, so the accessibility should allow anyone: from end-users to data protection authorities, to see all key information about privacy. We fulfill this criterion the same way that ads.txt does: a standard location and file name in the website directory. In this manner, anyone can find the websites' GDPR policies and it is always accessible when the website itself is online. This standardisation of location causes not only accessibility, it also increases machine-readability and consistency.

Machine Readability and Consistency

Privacy attributes can be significantly easier to audit when it is placed in a consistent manner, and machine readable format. This gives data protection authorities the opportunity to audit on a large scale: automated scripts can be used to audit a large number of websites at a time. Furthermore, data can be easily aggregated when the format is machine readable, which facilitates privacy research and large-scale auditing. Gdpr.txt uses a specified format that facilitates machine-readability. Using standardised lines in a text file that commit to the same format, while also facilitating comments for readability of the user. The consistency of such a file will also leave little space for omitting privacy attributes and give users a clear view of each of the most important aspects of how their data is collected.

Field	Name	Description
Field #1	Cookie name	The name of the cookie. This identifies which cookie is set. The website uses this together with the value to identify the cookie.
Field #2	Domain name of the cookie	The domain attribute of a cookie specifies which domain may receive the cookie. If this is the same as the host domain, that means it is a first party cookie.
Field #3	Duration of the cookie	The duration attribute specifies for how long the cookie is stored on the user's device. This is in the form of the number of days the cookies will remain on the user's device before it is expired and deleted.
Field #4	First or Third party cookie	This is a boolean attribute that indicates whether the cookie is a third party cookie. Thus means that the target domain is different from the host domain. It is placed on the website by someone other than the owner and collects data for that third party.
Field #5	Optional cookie	This is a boolean attribute which indicates whether this is an optional cookie or not. Optional cookies can be refused by the user, using the consent banner. When cookies are not optional they will always be placed on the user's device when they access the website, with or without consent.
Field #6	Http Only	This is a boolean attribute which indicates whether the httpOnly flag is set. This means that the cookie can only be transferred via HTTP, and therefore the cookie can only be accessed by the current server. This helps mitigate client-side scripts accessing the cookie data.
Field #7	Secure status	This is a boolean attribute which indicates whether the secure flag is set on the cookie. The secure flag causes the browser to only send the cookie over encrypted channels, therefore securing the communication between the user's device and the server.

Table 3.1: Record definition of cookie attributes in gdpr.txt files

http://example.com/gdpr.txt

```
# example.com/gdpr.txt
#
# Cookies
BIDUPSID, .example.com, 365, 0, 0, 1, 0
atpsida, .example.com, 0, 0, 0, 0, 1
NID, .google.com, 200, 1, 1, 1, 1
# Banner
example.com, 1
# Privacy policy
example.com/privacypolicy
```

Figure 3.2: Example of gdpr.txt file

Accountability

Eventually the standard should leave organisations accountable for the way they handle user data. It should be clear how the data is collected and to whom it is shared. While the gdpr.txt standard does not directly hold organisations accountable for their privacy decisions, it can be the means to an end. The data protection agencies have the basis in law to hold organisations accountable for misconduct of user privacy in the GDPR. However, this accountability is often not held, because as is shown by numerous studies that show lack of compliance on most websites. Accountability could be increased by using a transparency framework like gdpr.txt, by creating a more efficient way to hold organisations accountable.

Evolvability

Transparency frameworks should be evolvable in the sense that they should be able to adapt to new technologies, regulations, and other developments in the privacy field. In this manner, a framework can be used and evolved over time and there is no need to implement a completely new framework each time something in the industry changes. Gdpr.txt provides a simple grammar that can easily be adjusted as privacy regulations and concerns evolve over time. New lines and formats could be added, and existing ones can be easily updated while maintaining support of previous formats. Similarly, ads.txt added an update with new values that is adopted by the industry [iab tech lab, 2017].

Industry Support

As we have discussed in Section 2.7, a few of the privacy frameworks that exist or have existed decline due to lack of industry support. It is difficult to know what will catch on in actual organisations, and which methods are deemed too complex or simply impractical. However, as we have seen with the ads.txt and robots.txt standard, these frameworks can catch on in the industry. As we base our framework on ads.txt, which creates a self-disclosing and relatively easy to implement text file, we believe that this standard could equally get industry support.

Accuracy

Gdpr.txt facilitates accuracy in the sense that it is easy to update, and possible to automatically generate, as we will show in Section 4. The ease of updating creates the capability to always be up-to-date with any changes in the process of data collection and privacy. Furthermore, due to the machine-readability of the grammar, it can also easily be automatically created by machines, as we will demonstrate further. However, currently the accuracy of automatic generation does not include completeness, and the accuracy of files also requires manual updating by website developers. This could decrease overall accuracy of the standard, but these factors could also be mitigated by improved automatization.

3.3. Data collector

The goal of the data collector is to collect all data that is necessary for a gdpr.txt file in real-time. A visualisation of this component can be found in Figure 3.3. The tool consists of three components: the cookie collector, banner detection tool, and privacy policy lookup. These components will be discussed

in the following subsections.

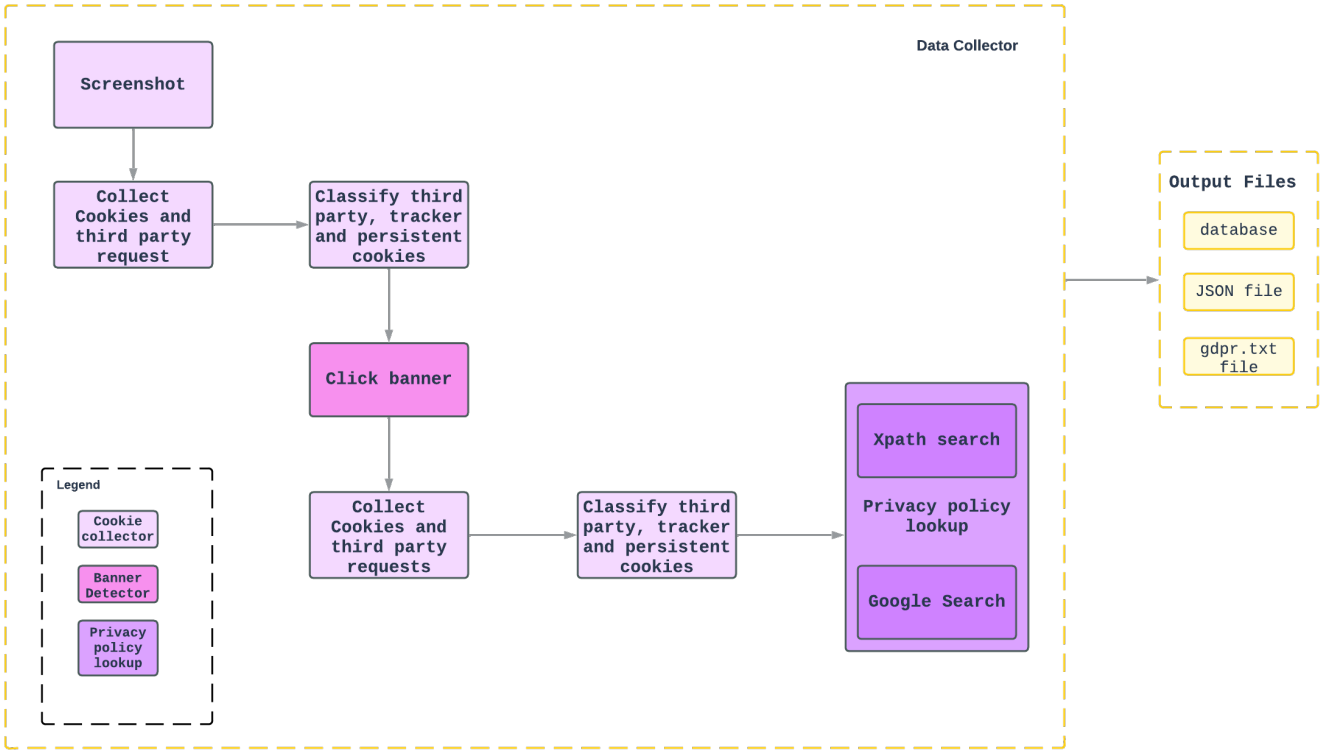


Figure 3.3: Data Collector Component of gdpr.txt tools

The data collector tool takes several arguments. Firstly, either a single website started with 'http', or a text file containing a URL on each line can be used as input. Multiple URL's can be analyzed in an asynchronised manner. Furthermore, the program takes the following arguments:

- `batch_size` Number of URLs to open simultaneously, default is 15
- `debug` Flag to log output for debugging
- `nd_json` Flag to store output as new line delimited JSON for use in e.g. BigQuery
- `screenshot` Flag to save screenshots
- `headless` Flag to hide actual browser windows
- `gdprtxt` Flag to create gdprtxt file and give file name
- `database` Name of database, default is gdpr.db

3.3.1. Cookie collector

The objective of the cookie collector is to crawl for real cookie data on any given website. For this we extend the auto-consent-check repository [de Wilde, 2022]. The description of the repository is the following:

> "Automatically check for GDPR/CCPA consent by running a Playwright headless browser to check for marketing and analytics scripts firing before and after consent. The software collects cookie data before and after consent and saves all the data to a JSON file."

To reproduce a visitor navigating through the different sites we chose to use the Playwright library [Microsoft, 2011] with Python. This allows us to have a complete browser that can be automatically controlled to do any action a normal user would. The user has the option to use headless mode, where no manual action can be performed. This means the browser is completely controlled from the Python

code and will not open an actual browser window. Playwright contains the function `cookies()` which returns all cookies [Microsoft, 2023]

Most of the attributes from the cookie can simply be read from the cookie text. However, it does not have an optional attribute. We determine the optional attribute by verifying whether the cookie appeared after the banner click, which would mean it activated after accepting all cookies. Furthermore we verify whether the cookie is persistent, and whether it is a tracker. A cookie is deemed persistent if it has a duration of more than 31 days, which means we evaluate the duration attribute against constant 31 to determine persistence. To classify tracker cookies, we make use of the most popular tracker domains list that can be found on the CookieCheck repository [Trevisan et al., 2019], the Justdomains list [Justdomains, 2022], and Disconnect.me blacklist [Inc, 2022]. All lists contain domain names and therefore a cookie domain is considered a tracker if it can be found in the list. Due to the fact that a cookie can be set on a subdomain at any depth level, we check up to the third level and the fully qualified domain name to determine whether it is a tracker or not. This means, if a subdomain is present we split on the dot and take only its last value, which is the third level. For example, in case of a.b.c.domain.com we will check a.b.c.domain.com, domain.com, and b.domain.com. The first one is the fully qualified domain name, the second one is the actual domain name, and then the last one is the third level. As can be seen, b.c.domain.com is not checked as it is at the fourth level.

The cookie crawler functions as following:

- Normalize URL with regular expressions
- Visit URL using Python Playwright
- Take screenshot (optionally)
- Capture third party requests pre-consent
- Capture cookies and attributes pre-consent
- Detect and click accept on consent banner with python Playwright
- Capture third party request post-consent
- Capture cookies and attributes post-consent

The result is a JSON file with all information, a gdpr.txt file according to the standards mentioned in section 3.2, a database file, and a short summary of statistics in the terminal. The gdpr.txt file and database file essentially contain the same information: the attributes from Table A.1, whether there is a banner and which CMP it is, and the link to the privacy policy. The JSON file is structured as can be seen in Figure 3.4. For all URL's, it includes all cookies and their attributes; all third party domains; third party domains that occur before consent; all third party cookies before consent; all tracking domains; tracking domains before consent; consent management provider name; the link to the privacy, including all links found; and the path to the screenshot.

Lastly, the structure of the summary is as following, where the summaries contain the mean, median and top 5 of all input websites:

- Website name and CMP
- Number of websites with tracker cookies
- Summary of tracker cookies per websites
- Summary of tracker domains per website
- Summary of persistent cookies per website

An example of a summary can be seen in Figure 3.5.

3.3.2. Banner Detection

Banner detection is implemented using the Playwright library and CSS selectors that commonly occur as accept buttons for consent banners. CMPs are valuable in banner detection, as they use the same structure on every website; a selector for one CMP banner will work for any website that implements that CMP. We implement banner detection per CMP, and also add standard banner keywords to find

```
[
  {
    "id": "aHR0cDovL2dkcHJ0eHQubmw=",
    "url": "http://gdprtxt.nl",
    "cookies_all": [
      {
        "name": "_koko_analytics_pages_viewed",
        "domain": "gdprtxt.nl",
        "expires_days": 0,
        "third_party": false,
        "httponly": false,
        "secure": false,
        "persistent": false,
        "tracker": false
      },
      {
        "name": "VISITOR_INFO1_LIVE",
        "domain": ".youtube.com",
        "expires_days": 180,
        "third_party": true,
        "httponly": true,
        "secure": true,
        "persistent": true,
        "tracker": true
      }
    ],
    "third_party_domains_all": [
      "https://abs-0.twimg.com",
      "https://jnn-pa.googleapis.com",
      "https://platform.twitter.com"
    ],
    "third_party_domains_no_consent": [
      "https://www.youtube.com",
      "https://abs-0.twimg.com",
      "https://platform.twitter.com"
    ],
    "third_party_cookies_no_consent": {},
    "tracking_domains_all": [
      "doubleclick.net",
      "google.com",
      "ggpht.com",
      "twimg.com",
      "googleapis.com",
      "gstatic.com"
    ],
    "tracking_domains_no_consent": [
      "youtube.com",
      "twimg.com",
      "twitter.com"
    ],
    "consent_manager": "cookie-yes",
    "privacy_policy": {
      "xpath_results": [],
      "google_results": ["https://gdprtxt.nl/privacy-policy/"],
      "link": "https://gdprtxt.nl/privacy-policy/"
    },
    "screenshot": "../screenshots/screenshot_aHR0cDovL2dkcHJ0eHQubmw=.png"
  }
]
```


Figure 3.4: Example JSON output file of gdpr.txt data collector tool. Summary of all cookies and their attributes, third party domains, tracking domains before and after consent; consent manager type; privacy policy link; and screenshot location.

Figure 3.5: Example summary output of gdpr.txt data collector tool. Summarizing the websites and their CMP type; for tracker cookies, tracker domains, and persistent cookies: mean, median and top 5 for consent and no consent.

custom banners, specifically their accept button.

For the implementation of the banner detection tool, we extended the auto-consent-check repository. The result includes all the most popular Consent Management Platforms that Nouwens et al. defined [Nouwens et al., 2022], including some more common CMPs. Furthermore, we added custom keywords manually by picking 10 random websites from the Netherlands database, and 10 random websites from the global database, and adding the keyword of the accept button if the button is not found yet. With this, a few keywords were added to the custom category, and the banner detection was made more effective. We verify the effectiveness of this tool in Section 4.3.1.

3.3.3. Privacy policy lookup

As mentioned in Section 2.5.2, it is recommended in the GDPR to have an independent page for its privacy policy that is easily accessible for users.

We applied two methods to determine whether a website has a privacy policy page, and where this page is located. In the first method, we take a JSON file which contains various keywords referring to privacy policies in multiple languages [Degeling, 2020]. The keywords are used to search in the source code of the website. The search is performed using an XPATH query that looks for anchor elements and returns the href link the elements point to. When we find a match, we store the link to the privacy documentation in a set object. After testing all keywords the object is converted to a list of unique privacy policy links.

If we do not find any link to the privacy policy using xpaths, we use a second method where we utilize Google search queries to find if there is a privacy documentation page available for the website. We created a function to which we pass a query to look for the privacy policy page on the specified website and returning a list of results. The query is made using the keyword 'site:', which filters all result to be exactly from the domain name we are interested in. The list of results is selected via JavaScript using a query selector that matches headings with the corresponding anchor element and returning the href link. To be sure that we do not include any url related to Google itself, we added a filter which would ignore the result if the title included 'googleadservices.com'. We also filter all links that do not contain the word privacy in their path. If the resulting list is not empty we set the privacy policy link to the first URL found, otherwise the error is reported and the link is marked as not found.

3.4. Parsers

We created two parsers to facilitate the auditing and ease of creating gdpr.txt files. One parser converts the gdpr.txt file into an sqlite3 database, and the other can convert databases to gdpr.txt files. The parser from database to gdpr.txt file is included in the data collector program to automatically create the gdpr.txt files. It uses regular expressions to normalize URLs.

3.5. Cookie Compare

The cookie compare tool takes two database files to compare each cookie entry. The component could be used for, among other, the following purposes:

- To analyze whether a gdpr.txt file is corresponding to the actual cookies in the webpage, and show if any cookies or cookie attributes differ.
- To analyze whether cookies change depending on changing variables. For example, in section 4.5 we do a controlled variable analysis for independent variables.

The tool is written in Python. As a cookie is exclusively identifiable by name and site domain, we compare these attributes match one cookie to another. Matching, in this case, means the identification of the same cookie in the two databases. When the cookie is matched, all cookie attributes are compared between the two, and if any difference is found in attributes, the values are saved to the JSON file. If a cookie is not matched, it is separately saved in the JSON file as well. The resulting file is formatted in the following manner:

- The matched cookies, per cookie, including unmatched cookie attributes and their values for both files.
- Unmatched cookies for the first file, by name and site domain.
- Unmatched cookies for the second value, by name and site domain.

4

Results

4.1. Description of Experiments

In the previous section we presented the tools for data collection and cookie matching , which in turn can be used to create gdpr.txt files and audit cookies across websites. In this section we demonstrate three experiments to evaluate the tools in popular browsers, and measure how accurate and practical they are for cookie registration and audit, while also getting an overview of the privacy landscape: specifically related to banners, CMPs and privacy policies. We show the effectiveness of banner lookup; privacy policy lookup; the percentages of banners, CMPs and privacy policies among different categories on the web; and how the cookie collection is affected by changing different variables.

4.2. Datasets

To perform testing and experiments on our solutions, we make use of three different databases containing popular websites: the Majestic Million dataset [Brown, 2021], Similarweb's Netherlands top website ranking [Similarweb, 2023b], and Similarweb's Top Website Ranking per category [Similarweb, 2023a].

The Majestic Million dataset is a collection of a million domains that find the most referring subnets, meaning websites are ranked according to how often they are referred to by other websites. This ranking gives an impression of popularity and 'importance' of websites. We use this website to show how the most popular websites that many people globally visit daily deal with privacy and show how our implementation can run on these websites. We use the top 500 of this database for our experiments, however, some websites might not always be reachable so we usually use between 480 and 500 websites.

Furthermore, we use two different types of Similarweb's datasets. Firstly, we use the top 50 most common Dutch websites to get an idea about privacy practises in a GDPR-governed country. This dataset is also used to test cookie collection in different circumstances. We use Similarweb's Top Website Ranking per category. This will be used to compare privacy practises among different categories of websites. Similarweb ranks websites based on user popularity: the ranking is based on unique visitors and pageviews on the main domain and all subdomains. The categorized datasets contain 50 domains per category.

4.3. Banner detection

For the banner detection experiment we aim to evaluate our banner detection tool, while also evaluating the prevalence of banners and CMPs among the most popular websites and popular websites within different categories.

4.3.1. Banner detection verification

The first part of the experiment consists of manual verification of the banner detection component. This experiment is performed by running the data collection tool on the top 50 websites in the Netherlands, and then manually verifying whether the cookie banner has been found, whether the banner actually exists, and if the program has managed to interact with the webpage and click the accept button. This is done by running the no-headless version of the program, and visually inspecting the clicking when it is done, while afterwards verifying that the correct result was recorded.

The results of this experiment can be found in table 4.1. As can be seen, the correct results was found 71 percent of the time, where either the banner existed and it was clicked, or the banner did not exist and it was not clicked. Considering we find the correct label for cookie banners around 71% of the time, the program can be used for automated generation of the gdpr.txt files, but should still be verified on whether it is found.

	Banner found and clicked	Banner not found
Banner	57.1%	10.2%
No Banner	16.3%	14.3%

Table 4.1: Banner verification top 50 NL

4.3.2. Consent Management Platforms

We use the CMP label to find out what percentages of analyzed websites use the most common Consent Management Platforms. CMPs are generally easier to detect than custom banners, as they all use the same structure and label for their banners. Getting an insight on the prevalence of CMPs gives us a better idea on how to perform banner detection. The results can be found in table 4.2. As can be seen, it is more common for websites to either have no banner (or not detected) or a custom banner. The CMPs are less common in the NL dataset, but worldwide One Trust seems to be the most used in popular websites. Furthermore, it is interesting to see that worldwide, about half of the websites deploy a cookie banner, while in the Netherlands, a GDPR country, the majority of 70% has a banner. This could be explained by the fact that countries outside the EU might bother less to implement these privacy features, or might not be aware of the regulations.

Moreover, it can be seen that the percentage of cookie banners worldwide, adding to 52% is quite close to what previous researchers have found, namely 60% for European websites, 44% for UK websites, and 48% for Greek websites as mentioned in section 2.6. We found a higher percentage of CMPs than found in the previous research of Bollinger et al. [Bollinger et al., 2022], who found only 3.5% of websites use CMPs and by Sanchez-Rola et al. who found 5% CMP usage on EU websites. The discrepancy in results could be explained by the fact that a larger dataset was used in the previous work, and the websites that are less popular and therefore more small scale, might employ less CMPs or cookie banners altogether.

Top 50 NL		Top 500 worldwide	
Custom	26	No Banner	235
No banner	15	Custom	108
One Trust	4	One Trust	82
TrustArc	3	TrustArc	8
Didomi	2	Quantcast	7
Percentage of websites utilizing Banner			
70%		52%	
Percentage of websites utilizing CMP			
14%		26%	

Table 4.2: Top 5 CMPs on top websites

4.3.3. Banner occurrence per category

In the next experiment we evaluate banner frequency among different categories of websites. The dataset we use for this is the Similarweb dataset that is indexed by category. The following categories were assessed: Arts and Entertainment, Business and Consumer, eCommerce and Shopping, Finance, Gambling, Health, Job and Career, and News and Media. These categories were chosen to give an accurate depiction of a variety of website categories and how they differ in privacy-related features.

We evaluate the occurrence of banners per category. The results can be found in Table 4.3. The average of the categories has a 30% occurrence of banners. The Gambling category stands out with a significantly lower cookie banner implementation rate of 3%. This suggests that the majority of entities within this category do not have cookie banners in place. The low implementation rate raises concerns about transparency and compliance with cookie consent regulations within the gambling industry.

The News and Media category exhibits a relatively higher cookie banner implementation rate of 48%, as well as the Health category with 45%. This suggests that a significant number of entities within this category have implemented cookie banners on their websites. The higher implementation rate indicates a proactive approach to inform users about data tracking practices and obtain their consent. It reflects the importance placed on user privacy and compliance with cookie consent regulations within the news and media industry, as well as the health industry. This could be an indication that a higher interest in privacy and compliance generally exists in the health and news industry.

Category	Cookie Banners
Arts and Entertainment	29% (12/41)
Business and Consumer	35% (13/37)
eCommerce and Shopping	22% (9/41)
Finance	30% (11/37)
Gambling	3% (1/40)
Health	45% (20/44)
Jobs and Career	31% (12/39)
News and Media	48% (19/40)
Average	30%

Table 4.3: Banner occurrence per website category

4.4. Privacy Policy Detection

The data collection tool was employed to automatically visit each of the top 500 websites. As mentioned in section 3.3.3, the tool uses Xpaths and Google search to find privacy policies for each website. The presence and location of privacy policies are recorded for each website.

To ensure the accuracy of the identified privacy policies, a manual verification was conducted on the top 50 websites in the Netherlands. Each website was examined on whether the privacy policy link works, directs to the correct domain, and contains the actual privacy policy. For the websites on which no privacy policy was found, we manually confirm whether it can be found. If it cannot be found manually, we consider it non-existent or unfeasible for end-users to find.

In the first experiment, we found no websites that do not have a privacy policy. Privacy policies were either correctly found, found but leading to the incorrect domain, or not found at all. The results can be seen in table 4.4. 80% of privacy policies were correctly found, while 16% recorded a link that did not lead to the correct page. 88% of these (7 out of 8) had privacy policies that redirected to a different domain. 63% (5 out of 8) of these were news-related websites. This is explained by two factors: news-related websites in the Netherlands often fall under the same larger organisation, and this organisation has a centralised privacy-policy page on a different domain; and news-related websites often have multiple URLs containing privacy related keywords, because they report on these sorts of topics. This is also an interesting result, as it shows that it is not uncommon for privacy policies to be

hosted on a completely different domain. This makes it significantly harder to automatically search for privacy policies on the web, as it is sometimes not even stated on the webpage for which domains the privacy policy applies to. The verification implies that privacy policies might not always be easy to find for users: if it cannot be found by xpaths, neither by a simple Google search referencing the website, the privacy policy would also be hard to find for users.

Correct privacy policy found	Incorrect privacy policy found	Privacy policy not found
80%	16%	4%

Table 4.4: Privacy policy verification top 50 NL

In the second experiment we detect how many privacy policies can be found for the top 500 websites worldwide. Using the data collection tool, we find privacy policies on 93% of the websites. The discrepancy between the verification set and the global set can be explained by the fact that the verification set contained a large number of news websites (9 / 50) which are harder to detect by the automated Google search.

As was mentioned in section 2.5.2, previous studies from 2019 and before have found 70 to 85% of privacy policies on popular websites. We find 93%, which could give the indication that the number of privacy policies have increased between 2019 and 2023.

4.4.1. Privacy Policy occurrence per category

We evaluate the occurrence of privacy policies per category. The results can be found in Table 4.5, where we find the average percentage of websites where privacy policies are found to be 80%. Similarly to the results of banner occurrence, the Gambling category stands out with a significantly lower compliance rate of 22%. This suggests that a large number of entities in this sector do not have privacy policies in place. Furthermore, the News and Media category stands out with a privacy policy occurrence of 98%. This suggests that a significant number of entities within this category have implemented privacy policies on their websites, but it could also be explained by the phenomenon found in the verification set: news websites are more likely to have the word 'privacy' in URLs of articles and might be more likely to host their privacy policy on a different domain, so this category might contain false positives. The Business and Consumer, eCommerce and Shopping, and Finance category also have a large privacy policy occurrence, with all of them occurring in 92% of the websites.

Because we again find a large discrepancy for privacy measures in the Gambling industry. The results imply that there should be more focus on privacy features in Gambling organisations generally.

Category	Privacy Policies
Arts and Entertainment	78% (39/50)
Business and Consumer	92% (46/50)
eCommerce and Shopping	92% (46/50)
Finance	92% (46/50)
Gambling	22% (11/50)
Health	82% (41/50)
Jobs and Career	82% (41/50)
News and Media	98% (49/50)
Average	80%

Table 4.5: Privacy Policy Occurrence per Website Category

4.5. Controlled-variable cookie analysis

We assess the practicality of our tools using three changing variables: Browser, Operating System, and Location. For these changing variables we measure which of the matched cookies' attributes are different depending on the variable.

In the controlled variable study we test the cookie collection component and the cookie compare component with a subset of random websites. For this experiment, we use the top NL dataset. Multiple experiments are done on different variables, with the following setup:

- We run the cookie collection tool on the subset with the baseline settings.
- We run the cookie collection tool on the subset with one variable change.
- The resulting databases are compared using the cookie compare component.

The baseline settings are the following:

- Browser: Chromium
- Location: Leiden, The Netherlands
- OS: Ubuntu 22.04
- Device: Dell XPS 13
- Date: May 2023

This will result in a list of the differences in cookie attributes and unmatched cookies. From the different tests we can conclude how cookies differ across browser, location and operating system. This can give us an indication on how the gdpr.txt generation will function depending on the different variables.

The following analysis will be performed:

- Baseline analysis: in this setting, we will use the above-mentioned baseline settings in two different runs to check for any differences in cookies.
- Browser comparison: here we compare two browsers: Chromium and Firefox to determine whether and which cookie attributes might be different depending on which browser is used.
- Location-based analysis: the program will be run in baseline settings using a VPN at two other locations: US and Spain. This is to compare the results with a country that is not in the EU, and one that is also in the EU. This could show how the GDPR affects cookie usage and how the tools can be used across borders. The following specific locations are used:
  - New York City, New York 10570, United States of America
  - Madrid 28013, Spain
- Operating system comparison: to determine which cookies might be different on different operating systems, we use Ubuntu 22.04 and Windows 10 to compare.

During the experiment we found that not all cookies are recorded in every run. This can happen due to multiple factors, including that cookies might load later, or that the banner is not always found. In this experiment we will look at cookie attributes, rather than the completeness of the cookie record. Furthermore, we do not visit all websites of the top 50 in each experiment, in the first two experiment we visit 38 to 44 websites, while in the experiments that use a VPN, we visit 34 and 32 websites respectively, this has to do with the bot protection of websites that often does not grant access to the Data Collector, especially when there is a VPN involved.

4.5.1. Baseline

The results of the baseline analysis can be found in Table 4.6. We can see that some attributes of cookies did get recorded differently: three duration attributes and one optional attribute.

When looking into the duration attribute of the websites (amazon.nl and vi.nl) manually, we notice that the cookies do in fact have different duration variables: in these cases the cookie quickly starts with one duration attribute, and shortly after the website loads, the duration changes. An example of this

attribute	changed cookies
duration	3
optional	1

Table 4.6: Baseline analysis

on the Amazon website can be seen in Figure 4.1. The reason for this could be that some cookies first load in a default stage, and afterwards they change to their intended state. However, this could also be used maliciously to make it seem like cookies are not persistent, while later changing their duration to being persistent and eventually tracking the user without their knowledge.

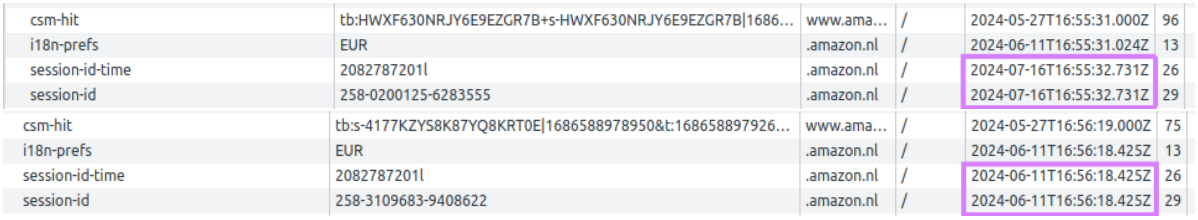


Figure 4.1: Example of change in cookie duration attribute

One website contains one cookie that has a different optional attribute. When manually inspecting this cookie, it turns out that the cookie is not optional. However, this could occur due to the delayed cookie loading on the website. We tag the cookie with the optional attribute under two conditions: the banner must be found, and the cookie should appear after the clicking of the banner has occurred. However, if a non-optional cookie loads after the banner click, it will therefore be marked as optional.

Aside from these small attribute changes, the websites all contain the same attributes for each cookie.

4.5.2. Browser comparison

The results of the browser comparison can be found in Table 4.7. As can be seen, a large number of duration and secure attributes changed, while some optional and http only attributes also differ between the results.

attribute	changed cookies
duration	84
optional	6
http only	2
secure	27

Table 4.7: Browser analysis

Firstly we will look at the duration attribute, where 12.2% of matched cookies differ. In 71 out of 84 cases, the duration of the cookie in Chromium was 400 days, while the duration of the cookie in Firefox was much longer, in 7 cases the duration was even 10 or 20 years. Notably, the websites on which this occurred were not malicious or suspicious websites, 5 of these being on two websites of the Netherlands' public broadcasting organisations (nos.nl and npostart.nl). The other 13 cookies with different durations among browsers were from a single website that had a duration of 30 days for Chromium cookie, and of 90 days for Firefox cookies. The duration difference between the browsers can be explained due to the fact that Chromium enforces a limit of 400 days on cookies [Chivukula, 2023], while Firefox does not seem to enforce a similar limit [Rajesh_Kumar_Yadav, 2022].

The difference in optional attribute can be explained similarly as the baseline: cookies that load after the banner click get labeled as optional. The reason why there is a slightly larger difference between the browser comparison vs the baseline is that the cookie loading times between browsers might differ. In all of the six cases we see that the Firefox cookie is marked optional, while the Chromium cookie is not. When inspecting manually we see that the cookies are in fact not optional. This could be an indication that cookies on Firefox will sometimes load later. An explanation of this could be that Firefox uses methods like Total Cookie Protection [Tim Huang and Edelstein,] to limit tracker cookies by default. The analysis by Firefox's protection features could results in longer loading times.

A single website also has two cookies that are http-only on Chromium, but not on Firefox. As this only occurs on one website we consider it an outlier.

Furthermore, we find 27 differences in the secure attribute: 26 where the Firefox cookie is not secure, while the Chromium cookie is, and one cookie where it is the other way around. Out of the 684 cookies that were matched in total, this makes almost 4% of cookies with unmatched secure attributes. Interestingly, 18 of these cookies are identifier cookies, which can uniquely identify a user and might be used for tracking purposes [Englehardt et al., 2015]. Another 8 of the cookies are tracker cookies of another kinds. When inspecting domains of the cookies, we find that 23 of them are third-party cookies, and 4 are first-party cookies. Another interesting observation is that all third-party cookies originate from known tracker domains. When manually inspecting the websites, we see again that the cookies first load in a sort of default format, and then change attributes very quickly. An example with the secure attribute can be found in Figure 4.2. Here we see a session tracker cookie that loads with the secure tag, but quickly (within a second) changes to non-secure.

Name	Value	Domain	Path	Expires / Max-Age	Size	H..	Secure	SameSite	Partitio...	Priority
session_tracker	qrnfgkolorjlmnmrmo.0.1687518631787.Z0FBQUFBQmt...	.reddit.c...	/	Session	237		✓	None		Medium
datadome	7uMRqYFYlPv8CJfnVwdWae1GiN-bfCRn-hOXuveZVmC...	.reddit.c...	/	2024-06-22T11:10:...	136		✓	Lax		Medium
session	932fe66e9dd4b190097c1b4d59f289094a0bf9e3gAWVS...	www.re...	/	Session	159	✓	✓			Medium
USER	eyJwcmVmcyI6eyJnbG9iYWxUaGVtZSI6IlJFRERJVCIsIm...	.reddit.c...	/	2024-07-27T10:39:...	344					Medium
edgebucket	RSZ2yzoxNS0ihu5FVR	.reddit.c...	/	2024-07-27T10:39:...	28		✓			Medium
csv	2	.reddit.c...	/	2024-07-27T10:39:...	4		✓	None		Medium
token_v2	eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJzdWIiOiJleU...	.reddit.c...	/	2024-07-27T10:39:...	1784	✓	✓			Medium
loid	000000000e08aqpv7u.2.1687516771269.Z0FBQUFBQm...	.reddit.c...	/	2024-07-27T10:39:...	226		✓	None		Medium

Name	Value	Domain	Path	Expires / Max-Age	Size	H..	Secure	SameSite	Partitio...	Priority
datadome	7uMRqYFYlPv8CJfnVwdWae1GiN-bfCRn-hOXuveZVmC...	.reddit.c...	/	2024-06-22T11:10:...	136		✓	Lax		Medium
session	932fe66e9dd4b190097c1b4d59f289094a0bf9e3gAWVS...	www.re...	/	Session	159	✓	✓			Medium
USER	eyJwcmVmcyI6eyJnbG9iYWxUaGVtZSI6IlJFRERJVCIsIm...	.reddit.c...	/	2024-07-27T10:39:...	344					Medium
edgebucket	RSZ2yzoxNS0ihu5FVR	.reddit.c...	/	2024-07-27T10:39:...	28		✓			Medium
session_tracker	qdqknjdcpedmdpfpki.0.1687516788614.Z0FBQUFBQmt...	.reddit.c...	/	Session	237					Medium
csv	2	.reddit.c...	/	2024-07-27T10:39:...	4		✓	None		Medium
token_v2	eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJzdWIiOiJleU...	.reddit.c...	/	2024-07-27T10:39:...	1784	✓	✓			Medium
loid	000000000e08aqpv7u.2.1687516771269.Z0FBQUFBQm...	.reddit.c...	/	2024-07-27T10:39:...	226		✓	None		Medium

Figure 4.2: Example of change in cookie security attribute on identity cookie

4.5.3. Location-based analysis

For the location based analysis we run the cookie collector on two other locations and compare it to the baseline in the Netherlands. We choose the US, because it is not in the EU and therefore the GDPR does not apply to its citizens. However, 40% of the websites in the top 50 NL dataset are American-based. With this analysis we could see if companies treat the users based in their country differently than those based in a EU-governed country. We use Spain as the other location to see how cookies differ between EU-member states.

United States

The results of the location based analysis compared to the United States can be found in Table 4.8. The changed attributes are 5 cookies in duration, and 5 in the optional attribute.

As for the duration, we find that all five cookies have a slightly lower duration (around 5 days) in the Netherlands, while in the US they have a duration of exactly 400 days. Four of these cookies reside on the same website (Twitter). As this is an international website, this could simply be because it is

attribute	changed cookies
duration	5
optional	5

Table 4.8: United States - Netherlands analysis

managed by different developers in both countries.

In all cases where the optional attribute differs, it labeled not optional in the US, while being labeled optional in the Netherlands. On manual inspection, the cookie is in fact optional in the Netherlands, while it is not in the US. It occurs on three international social media websites, which is most likely the case because they use different cookie policies in the US and Europe.

We can conclude, as concluded by previous research [Hu and Sastry, 2019], that the cookie usage in the US does not differ much from that in the EU. Some small differences in attributes occur, but there are no drastic differences in existing cookies.

Spain

In the comparison between the baseline datasbase as Spain, we find no difference in attributes. This is expected, as Spain mostly has the same laws for cookies as the Netherlands (the GDPR).

OS comparison

We compare cookies on Linux and Windows, the results can be found in 4.9. As can be seen, only duration attributes change on 4 cookies. All of the instances of this relate to Amazon cookies (both first and third party). We have seen a similar change in the baseline comparison where, among others, an Amazon cookie changes quickly in duration. Here we notice a similar result, where 4 cookies have a quick duration change and are therefor documented with different duration attributes.

attribute	changed cookies
duration	4

Table 4.9: Operating system analysis

5 Discussion

5.1. Contributions

We define a new privacy transparency standard that could have great impact on the privacy landscape when adopted. In this section we will discuss how we contributed to existing literature.

5.1.1. Criteria for Transparency Frameworks

We identify a definition for privacy transparency frameworks, and identify the following criteria for an ideal privacy transparency framework:

- Accessiblity
- Machine-readability and Consistency
- Accountability
- Evolvability
- Industry Support
- Accuracy

These criteria could serve as a benchmark for evaluating existing privacy transparency frameworks and facilitate a deeper understanding of the effectiveness and impact of various privacy transparency initiatives. Furthermore, the criteria can be used in development of new frameworks, as a practical guideline.

5.1.2. Gdpr.txt Transparency Framework

Establishing a method transparency

Gdpr.txt offers a powerful solution to enhance transparency in the realm of data collection and usage. By providing clear and standardized information about the cookies employed on a website, users can rely on the information that is given to them in the file. This empowers individuals to exercise greater control over their personal information, fostering a climate of trust between users and website operators. The method is created keeping the successes and downsides of previous solutions in mind, adopting a similar structure as industry-supported frameworks like ads.txt and robots.txt. Gdpr.txt could similarly get industry-support and widened accountability if it is adopted like the other txt frameworks. Gdpr.txt could even be added as a mandatory privacy feature to websites, which would help implementation rate and eventually transparency and auditability of the web.

Simplifying Auditing Processes

Under the GDPR, organizations are obligated to maintain robust data protection measures and be able to demonstrate compliance. The self-disclosing standard streamlines the auditing process by making it easier to assess cookie usage on websites, as well as privacy policies and banner usage. With a standardized format for disclosing cookie-related information, regulators and auditors can efficiently evaluate compliance levels, while reducing human effort and potential errors. This not only saves time but also ensures a more comprehensive evaluation of compliance across numerous websites. This

simplification of accessiblity and consistency also enables organizations to more effectively implement and maintain privacy controls, reducing the burden of compliance. The website developers can quickly understand how cookies should be implemented legally and technically, without having to dive into all the details of the GDPR and other privacy regulations.

Future Auditing and Transparency Software

One of the significant contributions of the gdpr.txt standard for cookie compliance is its machine-readable format. By adopting a structured and standardized approach, gdpr.txt allows for automated processing and analysis of cookie-related information. This machine-readable nature creates opportunities for future developments in auditing and transparency software. The standard enables auditors and regulatory bodies to leverage automated tools and algorithms for efficient and accurate auditing processes. If the gdpr.txt file were to be used widely, developers could create innovative transparency tools and applications. These tools can leverage the standardized data structure to provide users with enhanced insights into the types of cookies utilized, their purposes, and the associated data practices. By offering user-friendly interfaces, individuals could be empowered to navigate the complexities of online privacy with greater ease. This could provide more accuracy and evolvability than current privacy tools out there. These often scrape cookie real-time from websites, which can be unpredictable. Furthermore, gdpr.txt could pave the way for its scalability and adaptability to evolving technologies and regulatory requirements. As the digital landscape continues to evolve, new data protection regulations and technologies may emerge. The standard's machine-readable format ensures its compatibility with future auditing and transparency software, enabling seamless integration and flexibility.

5.1.3. Prototypes for Auditing Software

The study introduces multiple prototypes for software that can eventually be used for auditing of gdpr.txt files. The tools are meant to facilitate ease of auditing, and creating gdpr.txt files for more transparency. We introduce the following prototypes:

- Data collection tool
  - Banner detection tool
  - Privacy Policy detection tool.
- Cookie compare tool
- Parser tool

5.1.4. Insights into Privacy Landscape

We use the prototypes to gain insights into the current privacy landscape on the web. The study examines the prevalence of Consent Management Platforms (CMPs), cookie banners and privacy policies among analyzed websites in the Netherlands, global websites, and among different categories of websites. We show that privacy policy might not always be easy to locate.

Furthermore we show the difference in cookie attributes between browsers, location and operating systems on top websites from the Netherlands. Here we show that cookie attributes might not always be the same depending on different factors: duration is different among browsers, and some cookies change state quickly when a website is opened: we find this often in the secure attribute among id and tracker cookies.

These contributions provide insights into the prevalence of CMPs, cookie banners, and privacy policies, both globally and within specific website categories, shedding light on compliance with GDPR regulations and variations across different industries.

5.2. Limitations

5.2.1. Automated compliance and auditing limitations

The data collector could be valuable in collecting real-time cookie data for compliance purposes. However, it is subject to some limitations. Understanding these limitations will help users manage their

expectations and utilize the tool more effectively.

One notable limitation of the data collector tool is that it does not consistently collect all possible cookies of a website. Certain types of cookies, such as those activated by specific user interactions like mouse movement, may not be collected by the cookie collector tool. Additionally, certain cookies, such as those associated with third-party websites (e.g., YouTube cookies), may only become activated when the current browser visits those external sites. This is due to cookie syncing: where tracking data is exchanged between different websites [Englehardt and Narayanan, 2016]. We also show that cookie attributes can change depending on different factors such as the operating system, the browser that is used and the physical location. Consequently, the cookie collector tool may not always be able to collect all relevant cookies due to these activation dependencies.

Sometimes cookies can also load delayed or not at all due to other dependencies for cookie activation. Since the data collector tool uses one browser instance for the collection of the cookies, it is limited to this instance and will not collect any cookies that are not loaded at this time. This could be partly resolved by increasing the delay before collecting cookies. However, this would increase the run-time and a trade-off is introduced between completeness and run-time.

Furthermore we show that cookie attributes may change in websites, making them hard to document. We often find this in tracker and id cookies. For these cookies we cannot guarantee that all the attributes are properly documented, more specially for the secure attribute.

Another significant limitation of the data collector tool arises from the prevalence of bot protection mechanisms implemented by websites. Bot protection measures are designed to identify and block automated tools, including web agents, to safeguard the integrity and security of the website. Unfortunately, this can impede the smooth operation of the cookie collector tool, resulting in potential malfunctions and incomplete data collection. This is more likely to occur when a website is visited more repeatedly by the data collector. These protection systems aim to differentiate between human users and automated bots by analyzing various factors such as browsing patterns, interaction behavior, and IP addresses. Therefore the collector will sometimes not get access to the website and not be able to collect the relevant data. Furthermore, if the collecting agent is detected on the website, it could be possible to feed the tool false data. This could be employed by website who illegally track users, but want to appear compliant.

Users must be aware of these limitations, and consider that the resulting files will contain accurate cookie data, but not necessarily complete. When generating a gdpr.txt file with the automatic generator, the cookies should be manually investigated and checked as well. Similarly to auditing, the current prototype can show violations in cookie usage of e.g. persistent or tracker cookies, but cannot currently give a full report of all the deployed cookies. By understanding these challenges, users can adopt appropriate strategies to maximize the effectiveness of the tool and minimize data collection discrepancies.

5.2.2. Limitations of experiments

In the experiment, we used two main datasets. One of which has the 500 most linked websites worldwide, and the other uses top 50 of most visited websites in the Netherlands, and top 50 websites of different categories. These are relatively small datasets to conclude something about the whole web. The results give an idea of the prevalence of privacy features and cookie attributes on popular websites, but not the whole picture of the world wide web.

6 Conclusion

6.1. Summary

Many websites still do not comply to the General Data Protection Regulation (GDPR) after five years from its entry into application. This includes violations in the use of tracker cookies being used in 90% of highest traffic websites [Sanchez-Rola et al., 2019]; cookies are being set regardless of user consent [Matte et al., 2020]; cookie banners often deploy dark patterns to trick users into accepting cookies [Nouwens et al., 2020], among many other statistics that do not look favorable for GDPR implementation.

In order to secure European citizens' privacy, there should be more transparency on the web concerning cookie usage and user privacy. This will help users make more informed decisions, while creating more opportunities for efficient auditing of the GDPR and helping data controllers understand compliance better. We have addressed these issues by proposing a new transparency protocol: gdpr.txt.

Gdpr.txt is a self-disclosing, machine readable standard with a single reference point. It includes all cookies and their attributes, banner information, and privacy policy information. In order to evaluate privacy transparency frameworks, we set up the most important criteria for such a framework: accessibility; machine-readability and consistency; accountability; evolvability; and industry support. We show that compared to previous privacy transparency frameworks, the gdpr.txt standard scores well on all of these criteria. However, we do note that in the current implementation of automatic generation of gdpr.txt the accuracy, specifically the completeness, of the files cannot be guaranteed.

The study introduces multiple prototypes for software that can eventually be used for auditing and creation of gdpr.txt files, namely the following:

- Data collection tool: for real-time collection of cookies, privacy policy, and banner data, which can eventually be used to compare to gdpr.txt files, or create gdpr.txt files automatically.
  - Banner detection tool: designed to automatically accept all cookies in order to collect as many cookies as possible. The banner detection tool is 71% accurate in correctly identifying cookie banners and the corresponding accept buttons.
  - Privacy Policy detection tool: designed to find the link to the corresponding privacy policy of the website. The privacy policy detection tool correctly identifies 80% of privacy policy links.
- Cookie compare tool: for comparing of two cookie databases, which can eventually be used for auditing of gdpr.txt files against real-time cookies collected by the data collector, or comparison of two gdpr.txt files.
- Parser tool: parser for gdpr.txt file to database, or the other way around. This could simplify creation and comparison of gdpr.txt files.

When using these tools for creation of a gdpr.txt file, it should be manually verified by the website developer.

These tools are subsequently used to analyze the current privacy landscape on the web. We analyze a dataset of popular worldwide websites, as well as one with the most visited websites in the Netherlands, and a dataset divided by category.

Consent Management Platforms

The study analyzes the prevalence of various Consent Management Platforms (CMPs) among the dataset. The CMP label is used to identify the most commonly employed platforms. The results showcase that custom banners are more prevalent than CMPs. It is worth noting that the adoption of CMPs is less common in the Netherlands dataset compared to global usage, with One Trust being the most widely used CMP among popular websites worldwide.

We make an interesting observation regarding the presence of cookie banners worldwide. Worldwide 52% of websites deploy cookie banners, while in the Netherlands, an EU-based country, a majority of 70% of websites implement cookie banners. This discrepancy may be attributed to varying degrees of awareness and compliance with privacy regulations among different regions and industries.

Additionally, the study finds that the percentage of websites utilizing cookie banners worldwide (52%) aligns closely with previous research, namely 60% for European websites [Degeling et al., 2018], 44% for UK websites, and 48% for Greek websites [Kampanos and Shahandashti, 2021]. Notably, the research of Bollinger et al. reported a significantly lower usage rate of CMPs (3.5%) compared to the current study's findings (26%) [Bollinger et al., 2022]. This discrepancy may be attributed to the different dataset sizes used in the studies, with smaller-scale websites potentially employing fewer CMPs or cookie banners.

Privacy Policies

We utilize the data collection tool to automatically visit each of the top 500 websites. The presence and location of privacy policies were recorded for further analysis. This resulted in 93% of privacy policies being found. Comparing these findings with previous studies conducted before 2019, which reported privacy policy identification rates between 70% and 85% on popular websites [Liu and Arnett, 2002] [Nokhbeh Zaeem and Barber, 2017] [Degeling et al., 2018], the study's discovery of 93% suggests a potential increase in the number of privacy policies available between 2019 and 2023.

Banners and Privacy policies among Website Categories

We aim to gain insights into the prevalence of privacy-related features across commonly visited websites and identify any variations between categories, using a datasets of the most popular websites among different categories.

The findings reveal that the average occurrence of cookie banners across all analyzed categories is 30%. Notably, the Gambling category stands out with a significantly lower implementation rate of cookie banners, at only 3%. This suggests that a majority of entities within the gambling industry do not have cookie banners in place, raising concerns about transparency and compliance with cookie consent regulations within this sector.

In contrast, the News and Media category exhibits a relatively higher cookie banner implementation rate of 48%, followed closely by the Health category with 45%. These findings indicate that a significant number of entities within these categories have taken proactive measures to inform users about data tracking practices and obtain their consent. The higher implementation rate in these categories reflects the importance placed on user privacy and compliance with cookie consent regulations within the news and media industry, as well as the health industry. It can be attributed to the generally higher interest in privacy that exists in these sectors. Although we do note that there is still room for improvement.

Furthermore, the study also evaluates the occurrence of privacy policies per category. Table 4.5 presents the results, highlighting that the Gambling category stands out once again with a significantly

lower compliance rate of 22%. This suggests that a large number of entities within the gambling industry do not have privacy policies in place, indicating a need for increased focus on privacy measures within this sector.

On the other hand, the News and Media category stands out with a privacy policy occurrence of 98%, indicating that a significant number of entities within this category have implemented privacy policies on their websites. However, we acknowledge that this high occurrence could be influenced by the presence of the word 'privacy' in URLs of articles, potentially leading to false positive results in privacy policy detection. The Business and Consumer, eCommerce and Shopping, and Finance categories also exhibit a high occurrence of privacy policies, with all three categories having privacy policies on 92% of the websites. This again shows a higher interest for privacy in these sectors.

Cookie Attribute Comparison

In this study we find that cookies have different attributes depending on factors including type of browser, type of operating system, and physical location. We find that Firefox and Chromium mainly have different cookie expiry dates; this is because Chromium enforces a maximum cookie duration of 400 days while Firefox does not have similar measures. We also find that the duration of cookies can differ based on location: in the US we found five cookies with a different duration than in the Netherlands.

When comparing browsers we find that, regardless of browser, the secure attributes of cookies might change after the cookies load. This largely occurs in identity and tracker cookies. Resulting in the conclusion that the automated documentation of the cookies might be more difficult for these types of cookies, specifically identity and tracker cookies.

We conclude that location within the EU most likely does not make a difference for cookie attributes: when auditing the same websites from a VPN in Spain, we find all the same cookie attributes as in the Netherlands.

6.2. Future Work

6.2.1. Improve data collector tool for more complete data collection

One of the limitations is that the cookie collector prototype often captures an incomplete record of cookies. This could be partly mitigated by automating user interaction in the browser. For example, scrolling or clicking on the website will get some cookies activated. These user actions should be researched and tested to get a more accurate collection for cookies. This can involve carefully simulating user interactions, monitoring cookie activation dependencies, and refining the collection process accordingly. It could also include simultaneously browsing common third-party cookie website like YouTube, or Twitter to activate third party cookies through cookie syncing.

Another mitigating factor that could be implemented in future work is adaptive data collection to prevent bot protection activating when the program is running. This could include implementing measures that replicate human-like browsing behavior which may include introducing random delays between interactions, using multiple IP addresses, and adjusting browsing patterns to appear more natural.

Similarly to the cookie collector part, the banner and privacy policy collector could also be improved in future work. For the banner detector, more selectors and CMPs could be added to accurately find more banners and eventually record more cookies. The banner detection could also be improved by implementing a third-party banner detector. For example, consent-o-matic [Nouwens et al., 2022] includes 50 CMPs in their detection mechanism and would therefore make a good addition to the banner detection.

Lastly, the privacy policy detection component could be improved by e.g. improved privacy link filtering. To remove news articles about privacy, regular expressions might be used to limit the number of words in the URL. Furthermore, some practical study could be done on where websites generally

store privacy policies to improve the automatic search function.

6.2.2. Utilize data collection for more analysis or tools

The features implemented in the data collector are not completely utilized in our experiments. For example, we find the number of trackers per website, third party domains, and persistent cookies. These metrics can be used to evaluate the datasets we have used or even larger ones to get a more complete view of privacy regulation compliance on the web.

Furthermore, based on gdpr.txt and the introduced tools, other tools could be built that e.g. automatically check compliance of GDPR and point out which features could be improved. Website developers could audit and verify whether their websites are GDPR compliant, how many tracker cookies they use, which third party domains are used, etc. Eventually such a tool could also add to the gdpr.txt standard. For instance, it could be expanded by showing all third party domains, or trackers after consent is given.

6.2.3. Additional features

An add-on to browsers could be created which parses the gdpr.txt file into a more user-friendly format showing the types of cookies and trackers currently on the website that is visited. This could pop up when the website is visited to immediately inform the user better in their privacy.

The data collection tool could improve auditing capability by adding purpose classification per cookie. If the purpose of cookies would be classified, in e.g. necessary cookies, or advertising cookies, it could be easier for users and auditors to know what is happening to the data, and whether the cookies are compliant. It would also give website controllers a better idea of which cookies function in what manner, and eventually give website controllers more knowledge and agency over their compliance. This could be implemented using the CookieBlock method [Bollinger et al., 2022]. CookieBlock uses machine learning methods to classify cookies based on their attributes. As we have already collected attributes, CookieBlock can use this data to classify cookies and give a more complete perspective on cookie usage and compliance.

Gdpr.txt could increase transparency of user data in this age of data monetization and decreasing personal privacy. In this paper we aim to demonstrate a simple framework for privacy transparency and how this could be applied to better regulate user privacy on the web.

Bibliography

[Acar et al., 2014] Acar, G., Eubank, C., Englehardt, S., Juarez, M., Narayanan, A., and Diaz, C. (2014). The web never forgets: Persistent tracking mechanisms in the wild. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 674–689.

[Anderson, 2023] Anderson, B. (2023). Explained: Security.txt, humans.txt, ads.txt robots.txt.

[Assembly et al., 1948] Assembly, U. G. et al. (1948). Universal declaration of human rights. *UN General Assembly*, 302(2):14–25.

[Bashir et al., 2019] Bashir, M. A., Arshad, S., Kirda, E., Robertson, W., and Wilson, C. (2019). A longitudinal analysis of the ads. txt standard. In *Proceedings of the Internet Measurement Conference*, pages 294–307.

[Bollinger et al., 2022] Bollinger, D., Kubicek, K., Cotrini, C., and Basin, D. (2022). Automating cookie consent and gdpr violation detection. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association.

[Brown, 2021] Brown, K. W. (2021). majestic_million. `https://github.com/kyle-w-brown/majestic_million`.

[Chen et al., 2021] Chen, Q., Ilia, P., Polychronakis, M., and Kapravelos, A. (2021). Cookie swap party: Abusing first-party cookies for web tracking. In *Proceedings of the Web Conference 2021*, pages 2117–2129.

[Chivukula, 2023] Chivukula, A. (2023). Cookie expires and max-age attributes now have upper limit.

[Consulting,] Consulting, I. Art. 7 gdpr, conditions for consent, url =.

[Consulting, 2020a] Consulting, I. (2020a). Art. 13 gdpr information to be provided where personal data are collected from the data subject.

[Consulting, 2020b] Consulting, I. (2020b). Art. 4 gdpr, definitions.

[Consulting, 2020c] Consulting, I. (2020c). Art. 6 gdpr, lawfulness of processing.

[Consulting, 2020d] Consulting, I. (2020d). Gdpr, consent.

[Cranor et al., 2008] Cranor, L. F., Egelman, S., Sheng, S., McDonald, A. M., and Chowdhury, A. (2008). P3p deployment on websites. *Electronic Commerce Research and Applications*, 7(3):274–293.

[Dabrowski et al., 2019] Dabrowski, A., Merzdovnik, G., Ullrich, J., Sendera, G., and Weippl, E. (2019). Measuring cookies and web privacy in a post-gdpr world. In *International Conference on Passive and Active Network Measurement*, pages 258–270. Springer.

[de Wilde, 2022] de Wilde, D. (2022). auto-consent-checks. `https://github.com/dumkydewilde/auto-consent-checks`.

[Degeling, 2020] Degeling, M. (2020). Privacy wording. `https://github.com/RUB-SysSec/we-value-your-privacy/blob/master/privacy_wording.json`. [Online; Accessed: Apr. 3, 2022].

[Degeling et al., 2018] Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., and Holz, T. (2018). We value your privacy... now take some cookies: Measuring the gdpr's impact on web privacy. *arXiv preprint arXiv:1808.05096*.

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

[Englehardt et al., 2015] Englehardt, S., Reisman, D., Eubank, C., Zimmerman, P., Mayer, J., Narayanan, A., and Felten, E. W. (2015). Cookies that give you away: The surveillance implications of web tracking. In *Proceedings of the 24th International Conference on World Wide Web*, pages 289–299.

[Europe, 2021] Europe, I. (2021). Tcf – transparency consent framework.

[Findlay and Abdou,] Findlay, W. P. and Abdou, A. Characterizing the adoption of security. txt files.

[Fouad et al., 2018] Fouad, I., Bielova, N., Legout, A., and Sarafijanovic-Djukic, N. (2018). Tracking the pixels: Detecting web trackers via analyzing invisible pixels. *arXiv preprint arXiv:1812.01514*.

[GDPR.eu, 2020] GDPR.eu (2020). Art. 12 gdpr, transparent information, communication and modalities for the exercise of the rights of the data subject.

[Golla, 2017] Golla, S. J. (2017). Is data protection law growing teeth: The current lack of sanctions in data protection law and administrative fines under the gdpr. *J. Intell. Prop. Info. Tech. & Elec. Com. L.*, 8:70.

[Grimm and Rossnagel, 2000] Grimm, R. and Rossnagel, A. (2000). Can p3p help to protect privacy worldwide? In *Proceedings of the 2000 ACM workshops on Multimedia*, pages 157–160.

[Harding et al., 2001] Harding, W. T., Reed, A. J., and Gray, R. L. (2001). Cookies and web bugs: What they are and how they work together. In *The Privacy Papers*, pages 375–388. Auerbach Publications.

[Hausner and Gertz, 2021] Hausner, P. and Gertz, M. (2021). Dark patterns in the interaction with cookie banners. *arXiv preprint arXiv:2103.14956*.

[Hill, 2018] Hill, K. (2018). 'do not track,' the privacy tool used by millions of people, doesn't do anything. `https://gizmodo.com/do-not-track-the-privacy-tool-used-by-millions-of-peop-1828868324`. (accessed: 22.05.2023).

[Hu and Sastry, 2019] Hu, X. and Sastry, N. (2019). Characterising third party cookie usage in the eu after gdpr. In *Proceedings of the 10th ACM Conference on Web Science*, pages 137–141.

[Huth et al., 2018] Huth, D., Faber, A., and Matthes, F. (2018). Towards an understanding of stakeholders and dependencies in the eu gdpr. *Proceedings of the MKWI 2018*, pages 338–344.

[iab tech lab, 2017] iab tech lab (2017). ads.txt – authorized digital sellers.

[Inc, 2022] Inc, D. (2022). shavar-prod-lists. `https://github.com/disconnectme/shavar-prod-lists`.

[InteractiveAdvertisingBureau, 2018] InteractiveAdvertisingBureau (2018). Transparency and consent framework.

[Justdomains, 2022] Justdomains (2022). Domain-only filter lists. `https://github.com/justdomains/blocklists`.

[Kampanos and Shahandashti, 2021] Kampanos, G. and Shahandashti, S. F. (2021). Accept all: The landscape of cookie banners in greece and the uk. In *ICT Systems Security and Privacy Protection: 36th IFIP TC 11 International Conference, SEC 2021, Oslo, Norway, June 22–24, 2021, Proceedings*, pages 213–227. Springer.

[Kellett, 2021] Kellett, S. (2021). What is "do not track" (dnt) and does it work? `https://www.avast.com/c-what-is-do-not-track`. [Online; Accessed: May. 5, 2023].

[Koch, 2020] Koch, R. (2020). Cookies, the gdpr, and the eprivacy directive.

[Krisam et al., 2021] Krisam, C., Dietmann, H., Volkamer, M., and Kulyk, O. (2021). Dark patterns in the wild: Review of cookie disclaimer designs on top 500 german websites. In *Proceedings of the 2021 European Symposium on Usable Security*, pages 1–8.

[Liu and Arnett, 2002] Liu, C. and Arnett, K. P. (2002). Raising a red flag on global www privacy policies. *Journal of Computer Information Systems*, 43(1):117–127.

[Matte et al., 2020] Matte, C., Bielova, N., and Santos, C. (2020). Do cookie banners respect my choice?: Measuring legal compliance of banners from iab europe's transparency and consent framework. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 791–809. IEEE.

[Microsoft, 2011] Microsoft (2011). Playwright.

[Microsoft, 2023] Microsoft (2023). Browsercontext. `https://playwright.dev/docs/api/class-browsercontext`.

[Nokhbeh Zaeem and Barber, 2017] Nokhbeh Zaeem, R. and Barber, K. S. (2017). A study of web privacy policies across industries. *Journal of Information Privacy and Security*, 13(4):169–185.

[Nouwens et al., 2022] Nouwens, M., Bagge, R., Kristensen, J. B., and Klokmose, C. N. (2022). Consent-o-matic: Automatically answering consent pop-ups using adversarial interoperability. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7.

[Nouwens et al., 2020] Nouwens, M., Liccardi, I., Veale, M., Karger, D., and Kagal, L. (2020). Dark patterns after the gdpr: Scraping consent pop-ups and demonstrating their influence. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13.

[Persoonsgegevens, 2017] Persoonsgegevens, A. (2017). Cookies.

[Persoonsgegevens, 2019] Persoonsgegevens, A. (2019). Ap: veel websites vragen op onjuiste wijze toestemming voor plaatsen tracking cookies.

[Poullet, 2010] Poullet, Y. (2010). About the e-privacy directive: towards a third generation of data protection legislation? In *Data protection in a profiled world*, pages 3–30. Springer.

[Rajesh_Kumar_Yadav, 2022] Rajesh_Kumar_Yadav (2022). Cookie maximum lifespan.

[Reidenberg and Cranor, 2002] Reidenberg, J. R. and Cranor, L. F. (2002). Can user agents accurately represent privacy policies? *Available at SSRN 328860*.

[Sanchez-Rola et al., 2019] Sanchez-Rola, I., Dell'Amico, M., Kotzias, P., Balzarotti, D., Bilge, L., Vervier, P.-A., and Santos, I. (2019). 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*, pages 340–351.

[Schneier, 2015] Schneier, B. (2015). The hidden battles to collect your data and control your world. *Data and Goliath, London*.

[Similarweb, 2023a] Similarweb (2023a). Top websites ranking, all categories.

[Similarweb, 2023b] Similarweb (2023b). Top websites ranking, most visited websites in netherlands.

[Soghoian, 2011] Soghoian, C. (2011). The history of the do not track header.

[Solomos et al., 2019] Solomos, K., Ilia, P., Ioannidis, S., and Kourtellis, N. (2019). Clash of the trackers: Measuring the evolution of the online tracking ecosystem. *arXiv preprint arXiv:1907.12860*.

[Sun et al., 2007] Sun, Y., Zhuang, Z., and Giles, C. L. (2007). A large-scale study of robots. txt. In *Proceedings of the 16th international conference on World Wide Web*, pages 1123–1124.

[Tim Huang and Edelstein,] Tim Huang, J. H. and Edelstein, A. Firefox 86 introduces total cookie protection.

[Trevisan et al., 2019] Trevisan, M., Traverso, S., Bassi, E., and Mellia, M. (2019). 4 years of eu cookie law: Results and lessons learned. *Proc. Priv. Enhancing Technol.*, 2019(2):126–145.

[UserCentrics, 2012] UserCentrics (2012). Cookiebot webpage.

[Utz et al., 2019] Utz, C., Degeling, M., Fahl, S., Schaub, F., and Holz, T. (2019). (un) informed consent: Studying gdpr consent notices in the field. In *Proceedings of the 2019 acm sigsac conference on computer and communications security*, pages 973–990.

[W3C, 2001] W3C (2001). The platform for privacy preferences 1.1 (p3p1.1) specification.

[W3C, 2002] W3C (2002). P3p and privacy on the web faq. `w3.org/P3P/p3pfaq.html`. (accessed: 22.05.2023).

[Wheeler, 2012] Wheeler, E. (2012). 'how 'do not track' is poised to kill online growth. `https://www.cnet.com/tech/services-and-software/how-do-not-track-is-poised-to-kill-online-growth/`. (accessed: 22.05.2023).

[Wolff and Atallah, 2021] Wolff, J. and Atallah, N. (2021). Early gdpr penalties: Analysis of implementation and fines through may 2020. *Journal of Information Policy*, 11(1):63–103.

A

Gdpr.txt implementation guide

A.1. Background

gdpr.txt is a tool to aid the disclosure of data collection on the internet. It is a self-disclosing standard where all cookies get reported, along with banner and privacy policy information. This report is meant to show how to implement gdpr.txt to create transparency for data subjects and data protection authorities.

This project is based on the robots.txt, and ads.txt standard [1]. The key attribute is that the file is posted on the webserver, this inherently proves that the owner authored the file. This implementation guide was made with the help of the ads.txt implementation guide [2].

A.2. File format and location

The publisher of the ”/gdpr.txt” file must post it on the root domain and any necessary subdomains. The file should be accessible via HTTP and/or HTTPS under the standard relative path on the server host: ”/ads.txt”. The HTTPS request header should contain ”Content-Type: text/plain”, or Content-Type: text/plain; charset=utf-8” to signal UTF8 support.

The format consists of records, separated by line breaks. The records consist of the following form: -3cm0cm

$$< FIELD\#1 >, < FIELD\#2 >, < FIELD\#3 >, < FIELD\#4 >, < FIELD\#5 >, < FIELD\#6 >< FIELD\#7 >$$

or

$$< FIELD\#1 >, < FIELD\#2 >$$

or

$$< FIELD\#1 >$$

Furthermore, lines starting with # are considered comments, and therefor ignored. Lines containing # should be considered ignored by the data consumer from # on.

A.3. The Data Record

Table A.1 shows the contents of each field, considering the values of cookies.

[1] robots.txt: https://www.robotstxt.org/, ads.txt: https://iabtechlab.com/ads-txt/

[2] https://iabtechlab.com/wp-content/uploads/2022/04/Ads.txt-1.1-Implementation-Guide.pdf

Field	Name	Description
Field #1	Cookie name	The name of the cookie. This identifies which cookie is set. The website uses this together with the value to identify the cookie.
Field #2	Domain name of the cookie	The domain attribute of a cookie specifies which domain may receive the cookie. If this is the same as the host domain, that means it is a first party cookie.
Field #3	Duration of the cookie	The duration attribute cspecifies for how long the cookie is stored on the user's device. This is in the form of the amount of days the cookies will remain on the user's device before it is expired and deleted.
Field #4	First or Third party cookie	This is a boolean attribute that indicates whether the cookie is a third party cookie. Thus means that the target domain is different from the host domain. It is placed on the website by someone other than the owner and collects data for that third party.
Field #5	Optional cookie	This is a boolean attribute which indicates whether this is an optional cookie or not. Optional cookies can be refused by the user, using the consent banner. When cookies are not optional they will always be placed on the user's device when they access the website, with or without consent.
Field #6	Http Only	This is a boolean attribute which indicates whether the httpOnly flag is set. This means that the cookie can only be transferred via HTTP, and therefor the cookie can only be accessed by the current server. This helps mitigate client-side scripts accessing the cookie data.
Field #7	Secure status	This is a boolean attribute which indicates whether the secure flag is set on the cookie. The secure flag causes the browser to only send the cookie over encrypted channels, therefor securing the communication between the user's device and the server.

Table A.1: Record definition of cookies

A.4. Syntax Definition

The core syntax is a comma separated format, with six defined fields and one record per line. This means that records are separated by end of line markers.

Sequences of whitespaces or tabs will be ignored. Malformed data will also be ignored. No fields should contain tabs, commas or whitspace, and if they do they can be escaped with URL encoding.

A.4.1. Privacy policy and banner declaration records

Any record with one defined field will be marked as the privacy policy record. This contains one record of the location (URL) to the privacy policy of the domain.

Any record with two defined fields will be marked as the banner record. The banner records contains the URL of the visited website, and a boolean variable on whether there is a banner with consent options.

A.5. Example

Below is an example of a website where multiple cookies are set:

http://example.com/gdpr.txt

```
# example.com/gdpr.txt
#
# Cookies
BIDUPSID, .example.com, 365, 0, 0, 1, 0
atpsida, .example.com, 0, 0, 0, 0, 1
NID, .google.com, 200, 1, 1, 1, 1
# Banner
example.com, 1
# Privacy policy
example.com/privacypolicy
```

A.6. Implementation

The code used for automatic generation and auditing of gdpr.txt files can be found on `https://github.com/kokosnoob/gdpr.txt-tools/tree/master`.