# On Compliance of Cookie Purposes with the Purpose Specification Principle

Imane Fouad, Cristiana Santos, Feras Al Kassar, Nataliia Bielova, Stefano Calzavara

▶ **To cite this version:**

Imane Fouad, Cristiana Santos, Feras Al Kassar, Nataliia Bielova, Stefano Calzavara. On Compliance of Cookie Purposes with the Purpose Specification Principle. IWPE 2020 - International Workshop on Privacy Engineering, Sep 2020, Genova, Italy. pp.1-8. hal-02567022

**HAL Id: hal-02567022**
**https://inria.hal.science/hal-02567022**

Submitted on 7 May 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# On Compliance of Cookie Purposes with the Purpose Specification Principle

Imane Fouad
*Inria, France*
*imane.fouad@inria.fr*

Cristiana Santos
*Inria, France*
*cristianasantos@protonmail.com*

Feras Al Kassar
*SAP Labs, France*
*feras.al-kassar@sap.com*

Nataliia Bielova
*Inria, France*
*nataliia.bielova@inria.fr*

Stefano Calzavara
*Università Ca' Foscari, Italy*
*calzavara@dais.unive.it*

## Abstract

*The enforcement of the General Data Protection Regulation and the ePrivacy Directive relies upon auditing legal compliance of websites. Data controllers, as part of their accountability and transparency obligations, need to declare the purposes of cookies that they use in their websites. This leads to relevant questions such as: How should purposes be described according to the purpose specification principle? And how to ensure a scalable auditing, enabled by automated means, for legal compliance of cookie purposes?*

*In this paper, we investigate the legal compliance of purposes for 20,218 third-party cookies. Surprisingly, only 12.85% of third-party cookies have a corresponding cookie policy where a cookie is even mentioned. Overall, we find out that purposes declared in cookie policies do not comply with the purpose specification principle in 95% of cases in our automatized audit. Finally, we provide recommendations on standardized specification of purposes following the recent draft recommendation of the French Data Protection Authority (CNIL) on cookies.*

## 1. Introduction

Auditing legal compliance of websites within the EU Data Protection legal framework is of paramount importance. *Data Protection Authorities* (DPAs) are interested in making auditing as precise and scalable as possible to enable regulatory enforcement, and to react towards the expansion of complaints received since the General Data Protection Regulation (GDPR) [1] came into force in May 2018. *Data Protection Officers* (DPOs), who oversee and evaluate the overall compliance of the companies' websites, are also concerned in making the auditing scalable to ensure compliance.

While analysing the cookies present on a website, an auditor needs to capture the *purpose* of each cookie. This defined purpose can then help to determine whether processing is legally compliant, what safeguards the GDPR imposes, and which legal basis can be used. Ultimately, it is the *purpose* and the processing that must be used to determine whether or not a cookie can be exempted from consent [25]. Finally, only when it's declared which cookies require consent, one can verify whether a website is setting such cookies before any action of the user, and whether a cookie banner is compliant with the GDPR and with the ePrivacy Directive (ePD) [19], [6].

DPAs advocate that *all* cookies should – as a best practice – declare their purpose. The UK, Greek, Finnish and Belgian DPAs [41], [23], [26], [12] endorse as a good practice disclosure of clear information about the purposes of cookies, including strictly necessary ones. The guidance of the 29 Working Party (29WP) [7] notes that although some cookies may be exempted from consent, they are part of a data processing operation, therefore publishers still have to comply with the obligation to inform users about the usage of cookies prior to their setting.

In practice, we observe that some websites describe the purposes of cookies in the corresponding *privacy policies* (or in *cookie policies*). *But how are such purposes supposed to be defined?* Article 5(1)(b) of the

GDPR and the 29WP [5] elaborate on the *"Purpose Limitation"* principle. This principle mandates personal data to be collected (1) for specified, explicit and legitimate purposes only and (2) not further processed in a way incompatible with those purposes. In this work, we focus on the first component of this principle named *purpose specification*.

We first analyse the legal requirements of the purpose specification principle, and derive how cookie purposes should be described. With the aim of *automatic auditing of websites at scale* to ensure compliance, we then perform a large scale crawling: we collect 20,218 third-party cookies from 84,658 pages of the top 10,000 domains. Thereupon we search for cookie policies describing these cookies and extract their purposes to evaluate how many cookie purposes satisfy the legal requirements of purpose specification.

Our first result is concerning: only 12.85% of 20,218 third-party cookies have a corresponding cookie policy where a cookie is mentioned. Our second result exposes the illusion of the legal value of cookie policies: only 5% of cookies include a description of their purposes in well-structured tables. By processing such tables with automated means, we have extracted purposes for 997 third-party cookies out of 20,218 cookies collected in our experiment.

We conclude with guidelines to DPAs, DPOs and policy-makers to enable automatic auditing of websites. We substantiate that policy-makers should propose means to specify purposes in machine readable forms, and establish an ontology of purposes that comply with the legal requirements and reasoning under GDPR, ePD and other legal sources. For transparency and scientific purposes, we make available the dataset of 997 cookies and their purposes to the research community for further experiments [15].

## 2. Legal Requirements for Purposes

The following analysis on the legal requirements for purposes is based on the most authoritative legal documents in the domain of privacy and data protection law. In particular, we extract the arguments laid down in binding legal sources, such as the rulings of the Court of Justice of the EU (CJEU), and the legal rules laid down in legal provisions of the GDPR and the ePrivacy Directive (ePD). For a complementary analysis we resort to the non-binding guidelines by Data Protection Authorities (DPAs), 29WP and OECD.

**Availability**. The 29WP [9] [10] recommends that organisations should publish a privacy or cookie policy on their websites, wherefrom users are able to access necessary information on the purposes of cookies being used, including the ones of third parties. From this recommendation, we derive a first requirement of *availability* stating that the purposes of cookies should be available to users. The OECD Privacy guidelines [2] and the GDPR re-enforce the *predetermination* of purposes – they specify that before, and in any case not later than at the time of data collection, it should be possible to identify the purposes for which these data are to be used. The requirement of 'availability of purposes' stems also from the *transparency principle* (Article 5(1)(a) of the GDPR, and Recital 39 thereto) [10] which mandates an obligation of data controllers to *inform* the purposes of processing to the data subject (Article 13 (1)(c) and Recitals 58 and 60 of the GDPR). The CJEU ruling on Planet 49 [21] asserts transparency obligations about cookie purposes, which also hold for third parties with whom cookies are shared.

In the following, we describe which are the legal requirements to define purposes lawfully (demanded by Article 5 (1)(b) of the GDPR and the 29WP [5]). The *purpose specification* principle focuses on the initial purpose of data collection. It identifies three criteria for describing a purpose: *explicitness, specificity, and legitimacy*. We analyse and contextualize each requirement in the context of purposes for cookies.

**Explicitness**. The three following conditions must be met for a purpose to be explicit: i) Unambiguous: a purpose must be sufficiently unambiguous as to their meaning or intent; ii) Exposed: purposes need to be clearly expressed, revealed or explained. The 29WP [7] contends that it is not enough for information to be "available" somewhere in the website that the user visits; iii) Shared common understanding: the definition of the purposes must be understood in the same way by everybody. Criteria iii) could be measurable by user studies which are out of scope of this paper.

**Specificity**. Purposes should be precisely identified and clearly defined. Their formulation must be detailed enough to determine what kind of processing is and is not included within the specified purpose [5]. *Violations* occur when a purpose is too vague, general or overly legalistic. The 29WP [5], [10] give such examples: "improving users' experience"; "marketing purposes", "IT-security purposes"; "future research"; "we may use your personal data to develop new services and products"; "we may use your personal data to offer personalized services".

**Legitimacy**. Purposes should conform to a legal basis for processing and regarding cookies and tracking technologies, the eligible legal basis is consent (Article 5(3) ePD). In the context of cookies and cookie policies, this requirement of legitimacy is not directly applicable and therefore we do not study it in this paper, but scope it in our previous work [35].

**Discussion on explicitness.** Controllers can take *"appropriate measures"* [10] for providing information in view of fair and transparent processing in a *"easily accessible"* way. As such, we claim that the positioning of cookies in a table signifies best how cookie purposes can be "clearly expressed and revealed", based on three reasons: i) Cookie purposes are hard to find inside of a text. Previous works showed privacy policies are typically long, complex documents laden with legal jargon [42], [36]. Reading privacy policies for all the websites a user visits annually would take about 244 hours/year [30]. As a result, these policies are ineffective at informing relevant information like as purposes [37], [32], [14]. ii) Auditing purposes: we interpret legal requirements in terms of usefulness for auditing and compliance automated procedures. iii) Commonly sustained and recommended practice: presentation of structured information in a table format is recurrent, even if non mandatory, either by commonly visited websites (such as Google, Wikipedia, LinkedIn), and it is also recommended by the UK DPA [41]. The Belgian, French and UK DPA websites present cookie purposes inside of tables which include, for example: name, expiry date, content and purpose of cookies. Legal scholars, as Koops, [28] underline that both controllers and end-users will benefit if purposes are consistently specified in a table, or even in a machine-readable form to avoid data controllers to hide behind vague or very abstract-level purposes or to function creep into new, unspecified ones.

## 3. Extraction of Cookie Purposes

**Third-party cookies.** When a data subject visits a website, two types of cookies can be set in her browser: first and third-party cookies. *First-party* cookies are set in the user's browser by the site explicitly visited by the user or programmatically by the third party script included in the website (that however executes in the same "origin" of the visited website). When used in isolation, first-party cookies are capable to track users *only within one visited website*. *Third-party* cookies are set either (1) in the HTTP response by any third-party content (images, html files or even at the delivery of

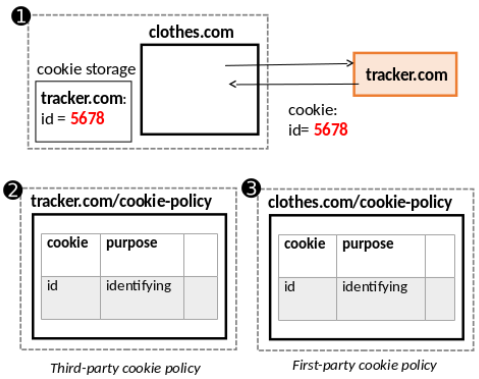


Figure 1: Third-party cookies and first- and third-party cookie policies.

scripts [24]); or (2) included via scripts operated from a third-party "origin" (a third-party origin most often is ensured by including a third-party iframe element, that includes a third-party webpage in the content of the visited website). Third-party cookies are capable to track users *across visited websites*.

In Figure 1, step ❶ demonstrates a hypothetical example of a visited website, `clothes.com`, a third-party cookie `id` set by a third party `tracker.com`.

We study only third-party cookies and their purpose descriptions for the following reasons: i) *third-party cookies are more likely to lead to privacy violations* [18]: by tracking users across websites, third parties can recreate a part of the user's browsing history which contains personal data. ii) *third-party cookies are usually not "strictly necessary" to the user visiting a website*: these cookies are usually related to a service that is distinct from the one that has been "explicitly requested" by the user [6]. As a consequence, third party persistent cookies are far more likely to require the user's consent. iii) *Third-party cookies allow third parties to track the user even if she has never visited the corresponding third-party server directly.* By storing a unique identifier inside a third-party cookie, third parties are able to recognise the user without having a direct interaction with her through a third-party server.

**Third party providers and legal responsibility.** Previous works have made large-scale measurements of the use of third-party cookies [42], [38]. However, the attribution of responsibility on the provision of information on purposes of third-party cookies was not explicitly determined yet. Since third party providers are *joint controllers* together with the first party website providers [8], Article 26 of the GDPR stipulates that *both* shall, in a transparent manner, determine their

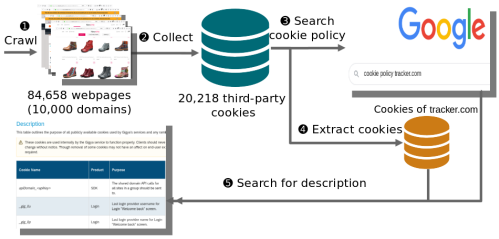


Figure 2: Overview of the data collection process.

respective responsibilities for compliance with information obligations (referred to in Articles 13 and 14) which include the purposes of the processing and their legal basis. As such, we argue that *third parties are also bounded to respect the principle of transparency and list both the cookies and their purposes on their own websites*. The CJEU also established that third parties need to provide information of cookies and their purposes in their own policies [20], [21], [22].

**Data collection.** Figure 2 summarizes all the steps of our data collection process. To collect third-party cookies for our experiment, we performed passive Web measurements using the Open Web Privacy Measurement (OpenWPM) platform [17]. While pretending to be a Web user, and maintaining the state of the browser, we automatically visited the top $10,000$ domains according to Alexa ranking [4] in February 2019 from a server located in France (step ❶ in Figure 2). For each domain, we visited the home page and the first 10 links pointing to pages in the same domain, resulting in data collection from 84,658 pages. We recorded cookies set both by Javascript and via HTTP Responses (step ❷ in Figure 2). We consider a cookie to be *a third-party cookie* if it's set by a different domain than the visited one. By domain, we refer to the $2^{nd}$-level TLD, such as `google.com`.

**Extraction of cookie policies.** For each domain that sets a third-party cookie at least once, we make a Google search of the cookie policy of the domain. For example, if a third-party cookie `id` from Figure 1 is set in the user's browser by `tracker.com`, we will search for the cookie policy of `tracker.com`.

To automatize the search, we search for the domain name concatenated with "cookie policy" in the Google search engine. For example, we search for *"tracker.com cookie policy"* to get `tracker.com`'s cookie policy (step ❸ in Figure 2). We then extract and store all the links $L$ from the first page of the resulting search results. We extract the subset $S$ of third-party cookies belonging to the same third-party domain from our crawling dataset. For example, we extract all the cookies set by `tracker.com` from all the pages we crawled (step ❹ in Figure 2). For each cookie in $S$, we search for the name of the cookie inside the rendered text of the extracted page of the cookie policy and save those where we found at least one mentioning of the cookie name (step ❺ in Figure 2).

For each third-party domain $d$ that owns a cookie in our dataset (such as `tracker.com` in our example), we extract a set of cookie policy links $L$. We define two types of cookie policies derived from the set $L$:

1) *Third-party cookie policy:* for each link $\ell$ in the set $L$, we first check whether it has the same top-level domain as $d$ or if they share the same *parents* organization. To extract the parent organization, we use the dataset built by Timothy Libert [39]. We call such link $\ell$ a *third-party cookie policy* because the cookie policy is directly provided by the owner of the third-party cookie. For example, step ❷ in Figure 1 shows the domain `tracker.com` that provides its third-party cookie policy tracker.com/cookie-policy with the list of cookies used by *tracker.com*.

2) *First-party cookie policy*: if no third-party cookie policy is found in the set $L$, we save all the cookie policy links that are hosted on other (first-party) websites. The cookie policy is hence provided by the first-party. For example, a cookie policy hosted on clothes.com/cookie-policy (see step ❸ in Figure 1) is a *first-party cookie policy* as it describes the cookie `id` set by `tracker.com`.

**Extraction of cookie purposes.** To extract purposes of cookies, we analysed first and third party policies separately because they need a different treatment.

We automatized cookie purpose extraction from *third-party cookie policies* using the following approach:

1) *The cookie name appears inside of a table:* We only consider tables because its representation is machine-readable and can be adapted to large scale studies.

2) *The length of the text does not exceed 1500 characters:* We use this criteria to discard tables not used for cookies descriptions, but rather used as the webpage representation style.

3) *The length of the cookie name is bigger than 1:* Single characters can be used inside a description as propositions (examples: *I*, *A*). Hence, we discard these cookies to reduce false positives in our results.

4) *The cookie only appears once inside of a table:* When the cookie name reappears several times inside

of a table, then either (1) the name of the cookie is a dictionary word in the language of the policy and so the description is not associated to a cookie, or (2) the cookie is referred in another cookie description, and in that case, we are not able to design which description defines the cookie purpose.

As to first-party cookie policies, we apply the same above approach and we further check that the domain name that set the cookie appears inside the description table as well. In fact, differently from the third-party policy directly provided by the cookie owner – where we are sure that the cookies in the description are those set by the third party– in case of first-party cookie policies, we need to check that both the domain that sets the cookie and the cookie name appear in the table. For example, in the first-party cookie policy of `clothes.com` in step ❸ of Figure 1, we search for the cookie name `id` together with the third party `tracker.com` that have set it.

**Limitations.** To extract cookie policies, we use google queries and we analyze the links from the first page of the resulting search results in an automated way, which enables a large scale study. However, our exhaustive cookie policies extraction methodology may not return policies following different pattern. To extract cookies purposes, we search for cookie names inside of cookie policies, and then extract the corresponding table row citing the cookie. When the cookie name belongs to the English dictionary, our exhaustive cookie description search algorithm may introduce some false positives. In such case, the description is using the English word and not providing a description of the giving cookie. We excluded all cookies with one character name to avoid introducing false positives. As a result, for these cookies we do not extract the cookie purpose descriptions even when they are available.

## 4. Evaluation of Cookie Purposes

In this section, we evaluate compliance of the extracted cookie purposes with the three requirements identified in Section 2. We explain the criteria adopted for each requirement and then we provide the analysis results. Notice that in this work we aim at *automated scalable auditing* and therefore we interpret legal requirements in terms of such auditing and compliance procedures. We thus take the position of a *website auditor*. Table 1 summarizes the results of this section.

**Criteria for availability**. We consider that a cookie is available if: i) a cookie policy exists; and ii) a cookie name is available in the cookie policy.

| | |
|---|---|
| **Total number of cookies** | 20,218 (100%) |
| **Cookies with available descriptions** | 2,598 (12,85%) |
| **Cookies with explicit descriptions** | 997 (5%) |

Table 1: Proportion of cookies compliant with the availability and explicitness requirements

**Results:** Out of $20,218$ third-party cookies, only $2,598$ ($12.85\%$) cookies satisfy the availability criteria: $423$ of them are mentioned in a third-party cookie policy and $2,175$ of them in a first-party cookie policy. In the following, we consider all the $2,598$ cookies.

**Criteria for explicitness**. As explained in Section 2, we suggest that a cookie purpose is explicit when described in a *structured table in the policy* because it is easier to identify the purpose for each specific cookie.

**Results:** Out of $2,598$ available cookies, only $997$ ($38.38\%$) cookies presented their description in an explicit way in a table (see Section 3 for details of our extraction algorithm). These $997$ cookies correspond to only $5\%$ of the total amount of $20,218$ third-party cookies we have collected demonstrating the illusion of the legal value of cookie policies.

**Criteria for specificity**. We consider that a cookie is specific if its description provides a clear and precise information about the purpose.

**Results**. We extracted $19,409$ cookie descriptions from first- and third-party policies of the $997$ cookies that have explicit purposes. Such high number derives from the fact that a single description can be repeated within first and third party policies. Out of $19,409$ cookie descriptions, $6,428$ are unique, however they describe $997$ cookies. This situation can be caused either by: i) the diversity of languages in cookie policies and the false positives introduced by our extraction algorithm; or ii) inherent confusion in the specification of purposes. Nevertheless, we observed that some cookies have different descriptions in different policies. Table 2 presents the 10 most popular cookie descriptions from a dataset of $19,409$ cookie descriptions. These top-10 descriptions occur $1,971$ times in our dataset, which constitutes $10\%$ of all the descriptions. Surprisingly, the top-5 descriptions of purposes do not render any specification about the use of cookies because the only statement provided for these cookies refers to their life span (session or persistent). The description conveyed in 6 seems to refer to 'Session Multimedia Content Player' cookies. Cookie 8 yields advertising purpose, for it refers to the collection of data with the purpose of optimizing ad display. Cookie 9 corresponds to the purpose of advertising to facilitate real-time-bidding.

| Row | Description | Occurrence | Specific |
|---|---|---|---|
| 1 | Pending Persistent HTML Local Storage | 365(1.88%) | ✗ |
| 2 | Pending Session Pixel Tracker | 267(1.38%) | ✗ |
| 3 | Pending Session HTTP Cookie | 233(1.20%) | ✗ |
| 4 | Pending 1 year HTTP Cookie | 220(1.13%) | ✗ |
| 5 | Purpose Expiry Type | 216(1.11%) | ✗ |
| 6 | Stores the users video player preferences using embedded YouTube video Session HTML Local Storage | 174(0.90%) | ✓ |
| 7 | Pending 1 day HTTP Cookie | 156(0.80%) | ✗ |
| 8 | Registers anonymised user data, such as IP address, geographical location, visited websites, and what ads the user has clicked, with the purpose of optimising ad display based on the users movement on websites that use the same ad network. 1 year HTTP Cookie | 125(0.64%) | ✓ |
| 9 | Used to present the visitor with relevant content and advertisement - The service is provided by third party advertisement hubs, which facilitate real-time bidding for advertisers. Session Pixel Tracker | 108(0.56%) | ✓ |
| 10 | Registers a unique ID that identifies a returning users device. The ID is used for targeted ads. 1 year HTTP Cookie | 107(0.55%) | ✓ |

Table 2: Top 10 cookies descriptions. *Occurrence: number of times the description is observed in a dataset of 19,409 cookie descriptions.*

Cookie 10 refers also to advertising, since the data collected is used for targeted ads.

## 5. Recommendations and Observations

Our experimental results confirm the common conjecture that cookie purposes are not described in a legally compliant way. In this section, we provide recommendations to policy-makers on how to improve the specification of purposes for trackers per requirement.

**How to improve specificity**. The top 5 cookie descriptions (see Table 2) show that purposes are rarely defined specifically. Purposes need to be pre-defined and modeled using ontologies that allow to reason about purposes inclusions, implications and generalisations. Such standardized approach would serve to minimize legal uncertainty [43]. Following our recent opinion [40] on the CNIL draft recommendation on cookies [25], *purposes should be defined in standardized taxonomies by the data protection authorities* to allow automatically reason about them.

The definition of purposes should be made with care because when users choose among many fine-grained purposes predefined in a system, they tend to opt for an open-ended "rest" category in which natural-language purpose descriptions are inserted [28].

**How to improve explicitness**. We found that for the 2,598 cookies that have cookie policies, cookie descriptions are often mixed with other text, which makes it hard to extract them. Only 997 cookies came with descriptions in well-structured tables. Following our opinion [40] on the CNIL draft recommendation [25], we propose that *each cookie should have only one standard purpose and a legal basis applied to it*.

Such standard description of each cookie and its representation in a table enables automatic large scale auditing of trackers. The same standard can be used in the design of cookie banners requesting users' consent.

**How to improve availability**. In Section 4 we found that only 2,598 (12.85%) out of 20,218 analysed cookies have a corresponding cookie policy wherein the cookie is mentioned. For the remaining 87.15% of cookies, no cookie policy was available. We suggest that cookie policies should be available on all websites to enable transparency of data processing purposes. We propose to use *a standard relative path on the server host, such as "/cookie-policy"* to enable its visibility. Similar self-declarative approaches are already used for websites: the declaration of access to crawlers in `robots.txt` file [34] and declaration of advertisers recently in `ads.txt` file [3].

## 6. Related Work

**Analysis of the purpose specification principle**. Basin et al. [11] analyse the purpose specification principle and propose a methodology for auditing GDPR compliance. Their analysis is limited to personal data collection in a business process context, while we analyse cookies that cover any kind of information (regardless of personal data) in the scope of web applications. Koops [28] analyses the purpose specification principle and, in an effort of techno-regulation, applies it within technical frameworks. The author suggests that purposes need to be specified, using a list of predefined domain specific purpose types. Grafenstein [43], [44] discusses this principle and propounds for a standardization of data purposes. We complement previous work by i) analysing the legal and theoretical framework on purpose specification for cookie purposes and denouncing their current ill-defined formulation; ii) considering the above mentioned proposals to mitigate the current state of the description of cookie purposes.

**Analyses of privacy policies**. To the best of our knowledge our work is the first to analyse purposes of

cookies within cookie policies. Reidenberg et al [33] considered several ambiguous and vague categories of privacy policies. Following Brodie et al [13], The Usable Privacy Policy project [46] combines technologies, such as crowd sourcing to develop browser plug-in technologies to automatically interpret policies for users. Ammar et al. [45] performed a pilot study, followed by a website privacy policy corpus [47]. This corpus have been later used by the Polisis tool that automatically extracts information flows described in privacy policies [27]. Morel and Pardo [31] made an extensive overview of privacy policies and tools used to analyse them at scale. Degling et al. [16] analyzed the availability of privacy policies on the top 500 websites before and after GDPR came in force. Libert [29] analyzed over 200,000 websites' privacy policies. In contrast to our work, Libert checked whether a website's privacy policy mentions transmissions to identified third-parties – he concluded that only 14.80% of such transmissions are disclosed.

## 7. Conclusion

In this paper, we assessed the scope of the principle of purpose specification and analysed whether it is respected in case of web browser cookies and their cookie policies. We found out that 95% of cookies do not have an explicitly declared purpose and hence are impossible to audit for compliance. The identified issues are rooted in the fact that data controllers have no explicit obligations to describe cookie purposes in a well-defined form. Policy-makers need to converge on harmonized requirements regarding the definition of purposes for cookies and other tracking technologies in the line with the 29WP guidelines [5]. DPAs and Standard Committees should standardize types of purposes for different contexts – this would minimize legal uncertainty, and reduce a case-by-case examination.

**Acknowledgements.** This work has been partially supported by the ANR JCJC project PrivaWeb (ANR-18-CE39-0008), the ANSWER project PIA FSN2 No. P159564-2661789/ DOS0060094 between Inria and Qwant, and by the Inria DATA4US Exploratory Action project.

## References

[1] "GDPR (EU) 2016/679, 27 April 2016."

[2] "OECD guidelines on the protection of privacy and transborder flows of personal data," 2013.

[3] "Ads.txt specification," https://iabtechlab.com/ads-txt/.

[4] "Alexa official website," https://www.alexa.com/.

[5] Article 29 Working Party, "Opinion 03/2013 on purpose limitation (WP203)."

[6] ——, "Opinion 04/2012 on Cookie Consent Exemption (WP 194)."

[7] ——, "Opinion 16/2011 on EASA/IAB Best Practice Recommendation on Online Behavioural Advertising."

[8] ——, "Opinion 2/2010 on online behavioural advertising, (wp 171)."

[9] ——, "Working Document 02/2013 providing guidance on obtaining consent for cookies', (WP208)."

[10] ——, " Guidelines on transparency under Regulation 2016/679, (WP260)," 2018.

[11] D. A. Basin, S. Debois, and T. T. Hildebrandt, "On purpose and by necessity: Compliance under the gdpr," in *International Conference on Financial Cryptography and Data Security*, 2018.

[12] Belgium DPA, "Guidance on cookies and other tracking technologies, 2020."

[13] C. A. Brodie, C.-M. Karat, and J. Karat, "An empirical study of natural language parsing of privacy policy rules using the sparcle policy workbench," in *SOUPS*, 2006.

[14] L. F. Cranor, "Necessary but not sufficient: Standardized mechanisms for privacy notice and choice," *JTHTL*, vol. 10, pp. 273–308, 2012.

[15] "Data," https://www.dropbox.com/sh/voi7levu2qgq9m3/AAC2SQ5iQ3Eu022BKK5HBwVla?dl=0.

[16] M. Degeling, C. Utz, C. Lentzsch, H. Hosseini, F. Schaub, and T. Holz, "We value your privacy ... now take some cookies: Measuring the GDPR's impact on web privacy," in *NDSS*, 2019.

[17] S. Englehardt and A. Narayanan, "Online tracking: A 1-million-site measurement and analysis," in *ACM CCS*, 2016.

[18] S. Englehardt, D. Reisman, C. Eubank, P. Zimmerman, J. R. Mayer, A. Narayanan, and E. W. Felten, "Cookies that give you away: The surveillance implications of web tracking," in *WWW*. ACM, 2015.

[19] "Directive 2009/136/ec of the european parliament and of the council of 25 november 2009 amending directive 2002/22/ec."

[20] European Court of Justice, "Case C-40/17 Fashion ID GmbH & Co.KG v Verbraucherzentrale NRW eV, ECLI:EU:C:2019:629."

[21] ——, "Case C-673/17 Verbraucherzentrale Bundesverband v. Planet49, ecli:eu:c:2019:801."

[22] ——, "Case C-210/16 Wirtschaftsakademie Schleswig-Holstein, ECLI:EU:C:2018:388."

[23] Finnish DPA, "Guidelines on confidential communications, 2020."

[24] I. Fouad, N. Bielova, A. Legout, and N. Sarafijanovic-Djukic, "Missed by filter lists: Detecting unknown third-party trackers with invisible pixels," in *PoPETs*, 2020, accepted for publication.

[25] French DPA, "Recommendation on "cookies and other trackers", 2020."

[26] Greek DPA, "Guidelines on cookies and trackers, 2020."

[27] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer, "Polisis: Automated analysis and presentation of privacy policies using deep learning," in *USENIX Security*, 2018.

[28] B.-J. Koops, "The (in) flexibility of techno-regulation and the case of purpose-binding," *Legisprudence*, vol. 5, no. 2, pp. 171–194, 2011.

[29] T. Libert, "An automated approach to auditing disclosure of third-party data collection in website privacy policies," in *WWW*, 2018.

[30] A. M. McDonald and L. F. Cranor, "The cost of reading privacy policies," *Journal of Law and Policy for the Information Society*, 2009.

[31] V. Morel and R. Pardo, "Three dimensions of privacy policies," working paper. [Online]. Available: https://hal.inria.fr/hal-02267641

[32] J. R. Reidenberg, N. C. Russell, A. J. Callen, S. Qasir, and T. B. Norton, "Privacy harms and the effectiveness of the notice and choice framework," in *Research Conference on Communication, Information and Internet Policy*, 2014.

[33] J. Reidenberg, J. Bhatia, and T. Breaux, "Automated comparisons of ambiguity in privacy policies and the impact of regulation," *J Legal Studies*, 2017.

[34] https://developers.google.com/search/reference/robots_txt.

[35] C. Santos, N. Bielova, and C. Matte, "Are cookie banners indeed compliant with the law? deciphering eu legal requirements on consent and technical means to verify compliance of cookie banners," https://arxiv.org/abs/1912.07144, 2019.

[36] C. Santos, A. Gangemi, and M. Alam, "Detecting and editing privacy policy pitfalls on the web," in *TERECOM@JURIX*, 2017.

[37] F. Schaub, R. Balebako, A. L. Durity, and L. F. Cranor, "A design space for effective privacy notices," in *SOUPS 2015*, 2015.

[38] J. K. Sørensen and S. Kosta, "Before and after GDPR: the changes in third party presence at public and private european websites," in *WWW*. ACM, 2019.

[39] "Timlib domains ownership," https://github.com/timlib/webXray_Domain_Owner_List.

[40] M. Toth, N. Bielova, C. Santos, V. Roca, and C. Matte, "Contribution to the public consultation on the CNIL's draft recommendation on "cookies and other trackers"," 2020.

[41] UK DPA, "Guidance on the rules on use of cookies and similar technologies', 2020."

[42] T. Urban, M. Degeling, T. Holz, and N. Pohlmann, "Beyond the front page: Measuring third party dynamics in the field," in *WWW*, 2020.

[43] M. von Grafenstein, *The Principle of Purpose Limitation in Data Protection Laws: The Risk-based Approach, Principles, and Private Standards as Elements for Regulating Innovation*, 1st ed. Nomos Verlagsgesellschaft mbH, 2018.

[44] ——, *Regulation as a Facilitator of Startup Innovation: The Purpose Limitation Principle and Data Privacy*, 2018.

[45] N. S. Waleed Ammar, Shomir Wilson and N. A. Smith, "Automatic categorization of privacy policies: A pilot study," Tech. Rep.

[46] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. G. Leon, M. S. Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell, T. B. Norton, E. H. Hovy, J. R. Reidenberg, and N. M. Sadeh, "The creation and analysis of a website privacy policy corpus," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1*, 2016.

[47] ——, "The creation and analysis of a website privacy policy corpus," in *ACL*, 2016.