

Me want cookie! Towards automated and transparent data governance on the Web

Wright, Jesse^{1,0}, Esteves, Beatriz^{2,0} and Zhao, Rui^{1,0}

¹Computer Science Department, University of Oxford, UK

²IDLab, Department of Electronics and Information Systems, Ghent University – imec, Ghent, Belgium

⁰These authors contributed equally to this work

Abstract

This paper presents a sociotechnical vision for managing personal data, including cookies, within Web browsers.

We first present our vision for a future of automated data governance on the Web, using policy languages to describe data terms of use, and having browsers act on behalf of users to enact policy-based controls. Then, we present an overview of the technical research required to *prove* that existing policy languages express a sufficient range of concepts for describing cookie policies on the Web today.

The authors view this work of automating privacy policies and consent in browsers as a stepping stone towards a future of semi-automated data governance at Web-scale, which in the long term will also be used by next generation Web technologies such as Web agents and Solid.

Keywords

Cookie, Browser, Data Terms of Use, ODRL, DPV, Negotiation, Data Governance, P3P, Do Not Track, Reasoning, Negotiation, Solid, Web, Agents

1. Introduction

In the ever-evolving landscape of digital privacy, the management of personal data, including cookies, within Web browsers has become increasingly crucial. Legislative attempts, to give users back control over data that is captured in cookies, have largely resulted in obstructive consent pop-ups across the Web containing long-winded policies which often are often deceiving or not upheld [1]. This has resulted in these terms of use being dubbed the “Biggest lie on the internet” [2].

In this paper, we present our vision of how ODRL [3], Data Terms of Use (DTou) [4] and DPV [5] can be embedded into websites with RDFa, and transmitted within HTTP Headers in order to well-describe, and allow negotiation over the terms of use that are applied to cookies in the browser. We argue that if deployed alongside regulatory incentives, or pressure, this technology stands to benefit individuals, industry and regulators. **Individuals** stand to benefit from a smoother experience and enhanced control over their privacy on the Web, with browsers managing data policies on their behalf. **Businesses** stand to gain significantly from this

NeXt-generation Data Governance workshop 2024, co-located with 20th SEMANTiCS, Amsterdam, Netherlands

✉ jesse.wright@cs.ox.ac.uk (W. Jesse); beatriz.esteves@ugent.be (E. Beatriz); rui.zhao@cs.ox.ac.uk (Z. Rui)

🌐 <https://www.cs.ox.ac.uk/people/jesse.wright/> (W. Jesse); <https://besteves4.github.io/> (E. Beatriz);

<https://www.cs.ox.ac.uk/people/rui.zhao/> (Z. Rui)

🆔 0000-0002-5771-988X (W. Jesse); 0000-0003-0259-7560 (E. Beatriz)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

standardized approach to cookie management. By adhering to machine-interpretable standards endorsed by regulators, companies can reduce the risk of privacy-related lawsuits, demonstrate compliance with regulations such as GDPR [6], the ePrivacy Directive [7] and CCPA [8], and streamline the implementation of privacy policies. Furthermore, it becomes easier for these bodies to implement automated auditing systems that validate their compliance with their advertised terms of use policies. **Regulators** stand to benefit from a unified framework for describing regulations around personal data in cookies, and automated techniques for checking compliance to that regulation.

The authors view this work as a critical stepping stone to achieve (semi-)automated data governance in data-centric technologies that form the next generation of the Web such as Web Agents [9] and Solid [10, 11, 12]. In particular, this is a problem space which can be immediately addressed with collaboration between academia, regulators and industry – thus enabling technologies such as ODRL [3], Data Terms of Use (DToU) [4] and DPV [5] to be ‘battle-tested’ and matured.

The article is organized as follows: Section 2 provides background information on Semantic Web technologies for the expression of policies and terms of use, Section 3 describes related work on privacy policies for browsers, vocabularies for expressing cookie preferences, and extensions for managing cookies, Section 4 describes our sociotechnical vision for managing personal data, including cookies, within Web browsers, Section 5 outlines interim technical-only approaches towards this vision, Section 6 presents an overview of how those terms-of-use languages introduced in Section 2 can be used to express the purpose descriptions of cookies, Section 7 presents an overview of in-progress work to assess the effectiveness of those languages introduced in Section 2 for describing cookie purposes – and evaluating the effectiveness of LLMs in generating terms-of-use descriptions from natural language, Section 8 presents call to action to a range of stakeholders to collaborate on working towards the vision outlined in Section 4 and Section 9 concludes the article with a discussion on future work.

2. Background

2.1. ODRL

ODRL is a W3C Recommendation for the expression of policies over digital assets. It includes a standardised information model [13] and a Vocabulary [3] to express flexible rules over data and services. This model allows the representation of permitted, prohibited and obligatory actions over assets, which can be further limited with constraints on rules, actions, assets and parties, and duties on permissions. The vocabulary can then be used to populate different types of policies with particular actions, functional roles of parties and specific constraints, e.g., temporal or spatial. Additionally, ODRL presents an extension mechanism, through ODRL profiles, which can be used to add further terms for specific use cases. However, a few shortcomings have been pointed out to the model, mainly founded on the lack of guidance over policy enforcement [14, 15]. As such, work on a formal semantics for ODRL is under-development, looking in particular at two scenarios related to access control and policy monitoring [16], with the goal of accurately describing the behaviour of an ODRL implementation. Active work on the representation of ODRL policies for personal data assets is also under-way [17, 18].

2.2. Data Terms of Use

Zhao and Zhao proposed the concept and a realization of perennial policy (a.k.a. Data Terms of Use, DToU), which targets at the challenges and utilities of decentralized Web, such as Solid [10]. In particular, one key goal is to support easy and smooth policy checking across applications and data providers, thus enabling users to make smooth and confident decisions on application authorization. Building upon principles of and addressing issues of existing policy languages, they proposed a novel policy language containing both data's (data provider's) policy and application's policy, and the reasoner supports checking the compliance between them as well as deriving data policy for output data. They have also demonstrated how the proposed reasoning engine is integrated with Solid.

2.3. DPV

DPV [5] is a community-based specification, being maintained and developed under the W3C umbrella, for the expression of metadata related to the processing of personal and non-personal data, based on legal requirements. DPV's main specification is a state of the art, jurisdiction-agnostic resource, containing meaningful taxonomies to describe entities, purposes, data and its processing, technical and organisational measures, legal bases, risks, rights and further privacy-related concepts. To invoke law-specific concepts, DPV 2.0 currently supports concepts from the EU General Data Protection Regulation (GDPR), the EU Data Governance Act (DGA), the EU AI Act, the EU Network Information Security Directive (NIS2), as well as extensions to specify personal data categories, location, risk management, technologies, justifications and AI terms [20]. Guidance documents for the adoption and usage of DPV [21] are also available, including guides for consent records, records of processing activities, data protection impact assessments and data breaches records.

2.4. RDFa

RDFa (Resource Description Framework in Attributes) [22] is a specification for enriching Web content with structured data, facilitating better data interoperability and search engine optimization. It allows Web developers to embed metadata within HTML, XHTML, and XML documents using standard HTML attributes like 'about', 'property', and 'content'. By annotating elements with RDFa attributes, developers can provide additional context and meaning to the content, making it more accessible to machines, such as search engines and other data processors, which can then extract and utilize this structured information for enhanced search results, richer snippets, and improved data connectivity across the Web.

2.5. Dark Patterns in Cookie Consent User Experience

Dark patterns in cookie consent popups manipulate users into agreeing to data collection. These tactics often violate GDPR and CCPA principles requiring informed and voluntary consent. Common dark patterns include: *Pre-selected Options* - Banners have pre-ticked boxes for non-essential cookies, contrary to GDPR guidelines requiring active consent; *Deceptive Button Colors* - Highlighting the "Accept" button more prominently than the "Reject" button to influence

user choice; *Complex Navigation* - Making users navigate multiple layers to reject cookies, while accepting them is straightforward [23]; *Misleading Labels* - Declaring marketing cookies as essential to imply users cannot opt out without affecting functionality [23]; *Hindering Withdrawal* - Making it difficult to withdraw consent by not providing easily accessible options; and *Manipulative Language* - Using vague or biased language to emphasize benefits of accepting cookies while downplaying data collection.

Studies indicate widespread use of these tactics, with only 11.8% of popular UK websites meeting GDPR consent requirements [24]. To comply with regulations and build user trust, websites should ensure equal prominence for “Accept” and “Reject” buttons, avoid pre-selected options, use clear language, and offer easy consent withdrawal.

3. Related Work

3.1. Privacy Policies in Browsers

The concept of privacy policies in browsers is not new. The World Wide Web Consortium’s (W3C) Platform for Privacy Preferences Project (P3P) [25] was a protocol published in 2002 aimed at “allowing websites to declare their intended use of information they collect about web browser users”. P3P enabled websites to communicate their privacy practices to users in a standardized and machine-readable format. P3P enables websites to encode their privacy policies in XML, which can then be automatically retrieved and interpreted by web browsers and other user agents. This allows users to easily understand a website’s data collection and usage practices without having to read lengthy privacy policies. Despite its innovative approach, P3P faced challenges in adoption and implementation, leading to limited use and eventual obsolescence as privacy concerns and regulations evolved [26].

Global Privacy Control [27, 28] and its’ “Do Not Track” (DNT) [29] predecessor take a sledgehammer approach to improving user privacy on the Web, introducing a binary signal which users can enable to indicate they don’t want their data to be sold or shared. Global Privacy Control (GPC) is an initiative designed to address some of the shortcomings of the “Do Not Track” (DNT) standard. GPC aims to provide users with a more robust and enforceable way to express their privacy preferences online. Unlike DNT, GPC has gained traction because it is backed by legal frameworks such as the California Consumer Privacy Act (CCPA) [30, 8]; although it is not yet enforced by the General Data Protection Regulation (GDPR) [31, 32]. These regulations require companies to honor user preferences for data privacy, giving GPC more legal weight [32]. GPC also benefits from clearer guidelines for implementation and stronger advocacy from privacy groups, making it a more promising tool for protecting user privacy in the complex digital landscape.

3.2. Vocabularies for expressing cookie preferences

The paper “What is in your cookie box? Explaining ingredients of web cookies with knowledge graphs” [33] introduces the *OntoCookie* ontology, developed specifically for this study to address the need for a machine-readable and standardised representation of Web cookies in compliance with the General Data Protection Regulation (GDPR). *OntoCookie* is a formal

representation of the cookie domain, comprising 229 axioms, 32 classes, 10 object properties, and 10 data properties. It models various types of cookies, their metadata, and their purposes (e.g., necessary, analytics, marketing) using a top-down ontology engineering approach. By leveraging this ontology, the authors created a knowledge graph (KG)-based tool to enhance user comprehension of cookie data sharing, providing a more transparent and interpretable view of cookie data.

Of the 32 classes introduced by OntoCookie, 8 are subclasses of Purpose (Analytics, Marketing, Profiling, ServiceOptimisation, ServicePersonalisation, ServiceProvision, Tracking), 10 are subclasses of Cookie (AuthenticationCookie, HostOnlyCookie, HttpOnlyCookie, PersistentCookie, SameSiteCookie, SecureCookie, SessionCookie, SuperCookie, TrackingCookie, ZombieCookie) and 2 are subclasses of Necessity (Necessary and Optional). With regards to the purpose classes introduced, only Tracking does not have an equivalent concept in DPV 2.0 (which otherwise has 88 additional purpose classes compared to OntoCookie), which may be worth adding as an extension to DPV under `dpv:Marketing`. The cookie classes are useful for adding additional descriptions about cookies, but do not help describe their *terms of use*, and are thus not as useful to this work.

Thus, we propose the use of ODRL / DToU & DPV in favour of OntoCookie as these vocabularies are better suited for describing terms of use in this use case, and better generalise to terms of use descriptions outside of the context of cookie policies which is the long-term end goal of this work.

3.3. Extensions for managing cookies

There have been several efforts to implement extensions which automate the process of accepting or rejecting cookies in browsers. For instance CookieBlock [34] is a browser extension designed to automate cookie consent management by using machine learning to classify cookies based on one of four purposes: Strictly Necessary, Functionality, Analytics, and Advertising/Tracking. The extension then automatically accepts or rejects cookies based on which of the four categories users have opted into. For classification, CookieBlock achieves a mean validation accuracy of 84.4% and filters out approximately 90% of privacy-invasive cookies without significantly affecting website functionality.

In addition to automating cookie consent, the study conducted by Bollinger et al. identifies widespread GDPR violations across nearly 30,000 websites, with 94.7% of these sites exhibiting at least one potential violation. The researchers highlight six novel types of violations, including incorrect category assignments and misleading expiration times. CookieBlock addresses these issues by enforcing GDPR compliance at the client side, autonomously detecting and mitigating violations. This approach does not depend on the cooperation of websites, ensuring that user privacy is protected even if the sites do not comply with GDPR requirements.

Thus CookieBlock, while innovative, acts as a bandaid solution by providing a temporary fix to the broader issue of cookie management and privacy. Since it relies on machine learning to classify cookies and automate consent, which, although helpful, does not address the root problem of ambiguous and complex privacy policies. Moreover, the fact that 94.7% of the sites that the authors studied did contain a potential GDPR violation underscores the need for a technical infrastructure, developed in collaboration with regulators, that facilitates platforms in

developing legally compliant terms of use.

By nature of the CookieBlock being ‘adversarial’ 10% of websites broke due to the cookies that were blocked. This issue would not arise in the long-term view of our proposal outlined in Section 4 as websites would always negotiate for valid cookie configurations. Our solution also offers more fine-grained control than CookieBlock where arbitrary descriptions of cookie purposes are permitted, rather than being confined to using 4 categories. Moreover, the solution we propose also allows terms-of-use to express a number of features other than purposes, such as retention period and recipient. In turn, users may also set preferences based on these features.

As a more minor point, in the CookieBlock style model - where clients are responsible for parsing the terms of use policies on websites - machine learning models are being called for each *user* visiting the website. In contrast, if the company itself is either (1) hand-crafting their machine readable policies using knowledge of how the company processes data internally or (2) using an LLM to translate human policies into a machine-readable form, clients only need to parse the machine readable policy which is much more computationally efficient.

4. Vision

We envision a future in which all websites present their cookie policies, and over time their privacy policies for *all* data processing, using formal languages such as ODRL or DToU in conjunction with vocabularies such as DPV.

In the near term, we propose that these descriptions be embedded in existing cookie popups using RDFa [35]. This allows the browser, or browser extensions, to directly parse the machine-readable policy and automatically accept/reject cookies based on user preference. Note that this also requires some standardised way for browsers to perform the action of accepting or rejecting cookies, such as standardised names for the checkboxes to opt in and out of cookies. Unlike existing browser extensions which accept/reject cookies [34], this solution does not rely upon LLMs or other heuristic measures to interpret the natural language cookie policies displayed, and accept/reject cookies based on broad categories of cookies such as *performance*, *functional* and *analytics*. Instead, this solution allows for matching against well-defined and fine-grained user preferences in the browser.

Similarly to [34], client side enforcement of some terms of use may be introduced once embedded RDFa descriptions are available, in particular, the browser may prevent the creation of cookies which may not be used for any purposes according to user preference. At this stage of the work this simply means that the only cookies permitted on a given website are those that have been ‘allowed’ by the extension. The extension can also enforce the *retention period* by deleting cookies once they have passed the expiry period.

Following this, we recommend migrating from websites having a binary notion of cookies being accepted or rejected, to cookies having a set of *terms of use*, which could be described in an agreement made between the website and the browser (on behalf of the user) encoded in ODRL or DToU. This, for instance, allows for a functionality where cookies may be accepted, but can only be used for a subset of the *purposes* or *actions* which a website had requested. In the long term, we propose that this agreement be added as metadata, in the form of an RDF dataset, to the

cookie itself. Since cookies are transferred to Websites via HTTP Request headers as specified in RFC 6265 we propose the addition of a 'Data-Policy' header which contains the user-agent specified terms-of-use policies for each cookie sent in the 'Cookie' header of a request. This proposal is backwards compatible with clients and browsers that do not support a Data-Policy headers as they may just ignore that header if they do not implement it. We propose the name 'Data-Policy' for the header in order to promote the description of additional data to cookies in the future. Under this paradigm, clients will still be able to enforce terms-of-use for those cookies that have no agreed upon usage purposes (which is equivalent to the cookie having been rejected). Clients will not be able to enforce terms-of-use agreements that allow the website owner to use personal data for some purposes. In this case, the user must 'trust' the Website owner to uphold the terms-of-use they have agreed to and declared in the 'Data-Policy' header. Note that this 'trust' may be supported by contractual agreements or regulation which will see the Website owner face penalties if they violate the terms-of-use that have been agreed to. As we shall re-iterate in the following subsections and Section 8, this is where a joint approach is required between research, regulatory bodies and industry; in order to develop a standard for this 'Data-Policy' headers which will:

1. Align with regulation, such that companies implementing policy engines to ensure their systems comply with the 'Data-Policy' attached to incoming cookies – receive certifications and a level of 'Safe Harbour' that guarantee they will not be subject to legal penalties. As a stricter approach, regulation could require the implementation of such policy engines and request regular audits from companies to demonstrate their compliance to the regulation.
2. Align with existing semi-automation approaches adopted for data governance within enterprises.
3. Incentivise industry adoption by promising the reduction of long-term engineering and legal costs to enterprises.

Looking further into the future, we expect the Website and browser to participate in a dialogue, most likely using HTTP headers or separate HTTP requests, in order to agree upon a set of policies to apply to cookies on that website. This dialogue facilitates sophisticated trade-offs, where users can negotiate the use of their data for marketing purposes, service optimization, or even offer micropayments for services instead of data sharing.

For the most part, we envision this negotiation consisting of a Website presenting a set of different conditions to be matched in order to be able to use the Website (e.g., to use the Website you opt in to use the set of Functional cookies for handling authentication), to receive a feed (e.g., on Facebook / Instagram) you must either opt into cookies for targeted advertising, have the user view a higher volume of advertisements, or have the user pay 1c for every 50 posts viewed. Observe that these are essentially requirements or terms of service that the Website advertises. The client would then compare these terms of service against the user's internal privacy preferences and then

1. Generate a data usage agreement to apply to each cookie in the browser
2. Perform any out-of-band operation required to complement the generated data-usage agreement; such as making a payment to the Website in lieu of opting into having data used for advertising purposes

We would like to explicitly note that this proposed part of the vision is highly speculative and dependent on jurisdictional data protection-related requirements. As previously discussed by Florea and Esteves, obtaining valid GDPR consent is still a challenging issue in decentralized settings, as it implies the user to know the purpose for which its data is being used for, as well as the identity of the legal entity ‘behind’ the processing of said data and a myriad of other conditions. As such, the development of a policy-based Web environment must not shy away from going beyond consent to explore other legal bases, while, of course, relying on it as an information safeguard for users.

This vision underscores a transformative shift towards a more user-centric, flexible, and legally aligned Web environment. In terms of user experience, there is a wide variety of ways in which browsers could allow user preferences to be set, including:

1. Users selecting a pre-defined configuration when first starting to use their browser.
2. Users manually configuring settings within an advanced settings menu.
3. Having the browser learn user preferences over time by presenting a popup to users when the negotiation requires knowledge of a user preference that has not yet been set.

Looking beyond cookies, the authors optimistically imagine a world in which all *private* or *personal* data sent over the Web is annotated with usage agreements between the sender and recipient – by extending the remit of the ‘Data-Policy’ header discussed above. In particular, we propose that both HTTP *requests* (where the client would be considered the sender and the server the recipient) and HTTP *responses* (where the client would be considered the recipient and the server the sender) be annotated with terms-of-use agreements encoded as RDF datasets. This enables the sender to be explicit about any requirements they have for how their data is governed, and for the recipient to implement policy engines that ensure these requirements are respected. In the worst case, if a server is not able to understand or handle the terms-of-use of an incoming HTTP request, we would expect the server to return a 5xx response and discard of any user data received in the request. It would be best practise for a server to also indicate any modifications required to the terms-of-use in order for a future request to be accepted, using an RDF-encoded response. If a client receives a set of terms-of-use that it is not able to respect, it should also immediately discard of the information received in this response.

These terms-of-use agreements would be applicable for data in the headers and message body with the possibility to granularly define different terms-of-use for different parts of the request or response. In particular, each Cookie in the Cookie header will have different terms-of-use, and these terms-of-use will differ from the terms-of-use applicable to the message body. A sensible place to begin implementing such terms-of-use agreements beyond cookies may be in Web Forms which send user data in the body of an HTTP POST request.

4.1. Beyond Websites

In the spirit of the Semantic Web [9, 37, 38], the authors posit that over time the Web will evolve away from users sending and receiving data via Websites, to having their interactions on the Web mediated via *Web agents* such as Charlie, the “AI that works for you.”; with most user information stored across personal data stores such as Solid Pods [10, 11, 12].

In this vision, we hypothesise that all data sent from personal data stores, and between personal agents will need to be annotated with terms-of-use to enable automated compliance with data governance requirements; as data is sent between agents representing data subjects with a range of preferences and legal rights, and hosted in personal data stores across a range of legal jurisdictions. These, and a range of other factors, will influence the terms-of-use that recipients must comply with when receiving data they wish to process.

Our hope is that by introducing a ‘Data-Policy’ header where agreed upon terms-of-use can be exchanged between clients and browsers; we prove the concept of terms-of-use agreements at Web scale; and this ‘Data-Policy’ header can be extended and re-used to support HTTP-based data sharing between Web agents, personal data stores and data processors.

4.2. Benefits to users

The proposed solution offers significant benefits to users by enhancing their control over personal data, improving privacy, and fostering trust in digital interactions. The current approach to cookie management, whereby users explicitly have to consent to the cookies they wish to have enabled on each website they visit, is inhibitive to everyday browsing. As a consequence, users are forced to choose between (1) reading through and interpreting extensive lists of cookie policies in order to manually accepting or rejecting the cookies a website has, (2) accepting or rejecting blanket lists of cookies such as functional, performance and analytical; with the UI to do so often only available after navigating UX antipatterns, or (3) give in to accepting all cookies in order to move to using the website as quickly as possible. In all 3 cases, a user’s experience of the Web is interrupted, resulting in a disrupted user experience.

The solution we propose allows users to set their preferences as browser settings, enabling automatic management of cookie preferences and eliminating the need for manual consent popups. This approach ensures that users’ privacy preferences are consistently applied without additional effort, thereby enhancing their control over personal data. Additionally, by selectively sharing data, users can continue to enjoy personalized experiences without compromising their privacy, allowing websites to offer tailored content and recommendations based on the data users are comfortable sharing. The system also provides transparency by clearly outlining data collection and usage practices, fostering greater trust as users feel confident that their data is handled responsibly and in accordance with their preferences.

4.3. Benefits to implementors

The implementation of standardized cookie management policies is anticipated to provide several benefits to businesses.

First, a “safe haven” provision could be established for companies that adhere to these standards, potentially reducing their risk of facing privacy-related lawsuits. By following the standard, businesses can demonstrate compliance and commitment to protecting user privacy, thereby mitigating legal risks [39]. In particular, regulators could offer automated compliance checks to verify that the privacy policies put in place by websites satisfy jurisdictional regulation such as GDPR and CCPA. In conjunction, companies would be able to implement internal auditing. By tagging all information incoming to their system with the applicable ODRL or

DToU policies, businesses can use policy evaluation engines to confirm that they are using personal data within the scope of *permission*, *prohibition* and *obligations* that have been granted.

Secondly, the ease of implementation is a significant advantage. Developers would find it simpler to generate privacy policies based on their system architecture, allowing for more accurate descriptions of data usage purposes. This streamlined process not only facilitates better compliance but also reduces costs. Legal teams would only need to review the selected policies, rather than drafting comprehensive terms-of-service documents from scratch. When policy changes (e.g., introducing a new functionality), the legal team can easily understand the difference from the comparison between old and new formal policy. In addition, they may have internal compliance tools built around the automated reasoning of such formal policies. This efficiency can lead to significant savings in both time and resources for businesses.

Furthermore, there are flow-on effects of the benefits provided to users. For instance, platforms are likely to have higher retention rates if users' are able to receive tailored online experiences without being concerned that their information is being used for purposes they are not comfortable with.

4.4. Benefit to regulators

Our proposal offers a number of benefits to regulators. With formal descriptions of cookie policies, there is a possibility of verifying whether website policies are coherent with local regulation in a semi-automated manner. This would likely reduce the cost of having legal experts do this process manually. Furthermore, the use of machine-readable terms-of-use agreements would improve the accuracy and efficiency with which regulators ensure that companies adhere to the agreements they make with users. This is because companies would be able to present system audits with standardized reports, which would be easier for regulators to review.

4.5. Comparison to related work on privacy policies in browsers

The core differences between our proposal and the related works have to do with the *expressivity* of the languages that we propose to use, and the differing legal context at the time in which we do our work. P3P [25] contains many similar concepts to those which we propose here, including having the ability to define cookie purposes from a fixed vocabulary and a trust engine to mediate between user preferences descriptions of cookie purposes.

From the perspective of expressivity, P3P proposes describing the following features of cookies Categories (What information is collected?), Purpose (How is it used?), Recipient (Who has access to it?), Retention (How long is it stored?) and Access (What information can the user access?). This is a subset of the concepts that are expressed by ODRL and DToU vocabularies as we show in later sections. Moreover P3P only offers 10 distinct purpose categories (and one 'other') category that are built into the specification and thus not extensible. In contrast, our proposal builds upon DPV for describing purposes which has 95 purposes at present, and is inherently extensible through OWL [40]. Furthermore, this extensibility ensures that semantic relationships are preserved. For example, one could define a new purpose, such as 'marketingShoes', as a subclass of the broader 'marketing' category to specify the exact nature of a cookie's function. However, user preferences related to marketing will still apply to this

new purpose through reasoning. Consequently, users who consistently opt out of marketing cookies will automatically be opted out of cookies designated for ‘marketingShoes’.

We recognise that there have been attempts to extend P3P with policies modelled in RDF [41] using Rei [42]. The primary goal of this work by Kolari et al. was to improve expressivity and adoption of P3P. The primary advantage of our proposed use of ODRL / DToU & DPV is (1) ODRL and DPV are more mature than Rei, (2) DPV has an extensive range of terms available for describing a wide array of privacy concepts, and (3) these vocabularies are built with modern regulation, such as GDPR [6], in mind.

Do Not Track (DNT) and Global Privacy Control (GPC) are less expressive by design, only offering a single signal to indicate that users do not wish to be tracked via cookies.

In terms of legal context, there is some level of consensus that P3P and Do Not Track were both unsuccessful due to a lack of legal pressures, however, as evidenced by the greater success of Global Privacy Control, there is a greater promise of adoption for such technologies when they either (1) are mandated in regulation such as the CCPA or GDPR or (2) offer companies more ‘safe haven’ from certain legal liabilities if adopted. This is why we do not propose an isolated technical solution, but rather a collaborative development between *research*, *industry* and *regulatory bodies* to work towards a sociotechnical solution by which the vocabularies used for formally describing *usage agreements* between the user and the website contain terms and concepts that have a well-understood legal interpretation.

5. Bandaid Solutions and Baby Steps

As an interim solution that requires client-side implementation only, and does not require involvement from regulators, website owners or CMPs – we propose a solution by which the ODRL and/or DToU terms-of-use descriptions for cookies are generated from their natural language descriptions using LLMs. The experiments we propose in Section 7 will determine the viability of this approach. These formal descriptions can then be used by browser extensions to automatically accept or reject cookies based on user preferences. This solution is a ‘bandaid’ solution in that it does not address the root problem of ambiguous and complex privacy policies, but rather provides a temporary fix to the problem of cookie management and privacy. However, it would improve the granularity of control that users have in comparison to existing solutions such as CookieBlock [34].

6. Describing cookies using ODRL and DToU

We now discuss how cookies policies can be described using ODRL and DToU respectively. In this paper we do not aim to prematurely conclude which of these vocabularies is best suited to describe cookie purposes and terms-of-use agreements between users and Websites. Instead, we provide an overview of the strength and weaknesses of the two languages for modelling cookie policies. In later sections we propose future work which we expect will shed light on

which language is most suitable in practise, and what modifications, if any, are required for the language to be adopted. In particular, we expect to be informed by:

1. performing experiments such as those outlined in Section 7 to test how well natural language cookie policies can be encoded in these formal languages;
2. co-design with regulators and industry as discussed in Section 8

As such, we shall present now how a cookie policy with the purpose description “Download certain Google Tools and save certain preferences, for example the number of search results per page or activation of the SafeSearch Filter. Adjusts the ads that appear in Google Search” is expressed using ODRL and DToU in Section 6.1 and Section 6.2, respectively.

6.1. ODRL

When it comes to the representation of cookie information as ODRL-based policies, such as that shown in Figure 1, two advantages can immediately be described in terms of flexibility and extensibility. In this context, ODRL has the flexibility to model distinct concepts embedded in the cookie description in human-readable language as machine-actionable elements, namely in terms of actions (processing operations) and purposes – this flexibility was already demonstrated for the particular use case of health data sharing by Pandit and Esteves [18]. Moreover, for the inclusion of new terms, ODRL provides extensibility through its profile mechanism – as shown by the usage of the ODRL profile for Access Control (OAC) which allows the expression of legally-aligned policies with DPV [17]. As for shortcomings, it should be pointed out that, by design, ODRL does not allow the expression of dynamic constraints, although a solution using property paths is being proposed to deal with this issue [15]. The resolution of this limitation is of particular importance for the modelling of temporal constraints and for cookies in specific when their retention period needs to be enforced or in case it needs to be updated. Furthermore, there are no direct matches for the ‘User Privacy and GDPR Rights Portals’ or ‘Wildcard match’ fields of the cookie.

6.2. DToU

On the other hand, DToU modelling features a standard mechanism to represent both the cookie policy (as app policy), such as that shown in Figure 2, and the user’s preferences (as data policy), as well as the formal semantics behind the modelling language. Thus, the standard reasoner can be directly used to verify their compliance without needing modifications. It provides an extensible *tag* mechanism (whose extension functions similar to an informal profile), which can be used to model the *purpose* constraints, as also demonstrated in [19]. Through enumeration and subclassing, the tag mechanism can be used to represent (discrete) temporal constraints – for example, by having longer durations as super-classes, and shorter durations as sub-classes of them, temporal constraints can be modelled as *requirement*, one of the two core types of tags. Having said that, a proper mechanism may be needed to represent temporal information without enumeration, particularly through extending the existing prohibition and

```

@prefix odr1: <http://www.w3.org/ns/odr1/2/>
@prefix dcterms: <http://purl.org/dc/terms/>
@prefix dpv: <https://w3id.org/dpv#>
@prefix oac: <https://w3id.org/oac#>
@prefix ex: <http://example.com>

<https://example.com/cookie-policy-grooveshark> a odr1:Request ;
  odr1:uid "8dc5d7e3-e31f-421a-8bad-6540172d787f" ;
  dcterms:description "Download certain Google Tools and save certain preferences, for
    example the number of search results per page or activation of the SafeSearch
    Filter. Adjusts the ads that appear in Google Search." ;
  dcterms:creator ex:google ;
  dcterms:issued "2024-06-03T17:58:31"^^xsd:dateTime ;
  odr1:profile oac: ;
  odr1:permission [
    odr1:assignee ex:google ;
    odr1:action oac:Download, oac:Store, oac:Profiling ;
    odr1:target <https://example.com/grooveshark-cookie-data> ;
    odr1:constraint [
      dcterms:title "Purpose for processing is to conduct marketing in relation to
        organisation or products or services." ;
      odr1:leftOperand oac:Purpose ;
      odr1:operator odr1:isA ;
      odr1:rightOperand dpv:Marketing ] ;
    odr1:constraint [
      dcterms:title "Rule can be exercised in the next 2 years." ;
      odr1:leftOperand odr1:elapsedTime ;
      odr1:operator odr1:eq ;
      odr1:rightOperand "P2Y"^^xsd:duration ]
  ] .

ex:google a dpv:DataController ;
  dpv:hasName "Google" ;
  foaf:page <google.com> .

```

Figure 1: A hand crafted ODRL description of the cookie policy

activation condition mechanisms. Furthermore, some informational fields are missing from the current DToU policy language, such as creator and creation time.

7. Analysing the effectiveness of ODRL and DToU with DPV for describing cookie policies

In this section, we propose a methodology that the authors of this paper are currently implementing in order to analyze the effectiveness of ODRL and DToU with DPV for expressing cookie policies. Our progress towards implementing this methodology is available at <https://github.com/jeswr/cookie-analysis/tree/chore/noise>.

```

@prefix dtou: <urn:dtou:core#>.
@prefix dpv: <https://w3id.org/dpv#>.
@prefix dur: <http://example.com/duration#>.
@prefix ex: <http://example.org/>.

ex:ap a dtou:AppPolicy;
    dtou:name <https://url-to.website/>;
    dtou:input_spec ex:cookie1 .
ex:cookie1 a dtou:InputSpec;
    dtou:data <https://example.com/grooveshark-cookie-data>;
    dtou:port [ dtou:name "google-cookies" ];
    dtou:purpose [ dtou:descriptor dpv:Marketing ];
    dtou:expect [ dtou:descriptor dpv:Download ],
                [ dtou:descriptor dpv:Store ],
                [ dtou:descriptor dpv:Profiling ];
    dtou:provide [ dtou:descriptor dur:two-year ];
    dtou:downstream [ dtou:app_name <https://google.com>; dtou:purpose dpv:Marketing ].

```

Figure 2: A hand crafted DToU description of the cookie policy

7.1. Dataset

For our analysis, we use a dataset of approximately 304,000 cookies which Bollinger et al. collected to perform their work on CookieBlock. These cookies were collected from around 30,000 websites that use Consent Management Platforms (CMPs).

The researchers selected CMPs that list cookies with their purposes. They then extracted cookies declared by the CMPs and those created during interactions with the websites. This process resulted in a comprehensive dataset of declared and observed cookies. The dataset includes around 304,000 cookies, with details such as their names, domains, expiration times, purposes and categories.

7.2. Research Challenge

For most properties, we are able to define a static rules-based mapping from this schema into ODRL and DToU descriptions similar to those shown in Figure 1 and Figure 2. The key property which requires linguistic interpretation is the Purpose. In particular, the dataset collected contains a natural language description of the purpose for which a cookie is being collected, whilst in ODRL and DToU we seek to map this to a list of well-defined DPV purposes to associate with the `odrl:constraint` and `dtou:purpose` properties respectively. Consequently, we seek to experimentally answer the research questions:

1. Does the current DPV Purpose vocabulary contain all concepts required to describe cookie purposes, if not, what new concepts are required.
2. Is it possible to accurately automate the mapping from natural language descriptions of purposes to DPV descriptions.

In the following sections we describe an experimental methodology which we are in the process of implementing in order to answer these research questions.

7.3. Methodology

We use a SPARQL [43] query engine to dereference the DPV ontology and query for the definition, label and note of all `dpv:Purposes` according to the SPARQL query found [here](#).

From this we generate a document listing all of the definitions, labels and notes with an anonymised ID. This document can be found [here](#). For each cookie we wish to classify, we pass this document to an LLM along with the name, category and description of the cookie. We prompt the LLM to identify the IDs of any relevant DPV purposes, if there is a part of the cookie description that is not captured by the current purposes, we ask the LLM to propose a description of a new DPV purpose that it would use in the description of the cookie purpose. The prompt used for this generation is available [here](#). A sample response from the LLM can be found [here](#). Note the explanation is requested to encourage the LLM to perform chain-of-thought reasoning when performing purpose classification; and this explanation is not used elsewhere.

We are in the process of analyzing the LLM proposed purposes to develop a new dataset of purposes to be proposed to DPV. The methodology we are planning to apply to achieve this is as follows:

1. For all of the new DPV purposes proposed from analysing the 304,000 available cookies we insert the purpose description into a vector database.
2. Group purpose descriptions by embedding similarity
3. For each group of purposes with high similarity, pass that set of descriptions back to the LLM and prompt it to:
 - a) Propose a purpose description that aligns with the input set
 - b) Propose a name for the new purpose
 - c) Identify whether the proposed purpose is a subclass or superclass of any of the existing purpose descriptions
 - d) Output this information according to a template `.ttl` document

7.4. Proposed Evaluation

We propose that the quality of the LLM classification and generation of purposes be assessed by a set of legal experts. For this evaluation a random sampling of the cookie dataset will be selected and legal experts will be provided with the cookie purpose descriptions, and asked to

1. select the set of DPV purposes they believe apply
2. describe any concepts they believe are missing

If both they and the LLM have identified that there are no DPV concepts to accurately describe the cookie purpose, the legal expert will then be asked to identify whether the new concept(s) proposed by the LLM match the concept(s) they propose.

8. Call to action

We call upon legal and policy experts working in the space of data governance to participate in co-designing machine-readable cookie description and transmission standards, and automated cookie selection. In particular, we call upon supervisory bodies, such as the European Data Protection Board (EDPB) or the Information Commissioner’s Office (ICO), that are capable of enforcing compliance with the standards we propose for terms-of-use descriptions, to ensure compatibility between the architectures we build, and the regulatory frameworks of the regions in which they will be deployed.

At the same time we call upon industry, including CMPs providers, to co-design a solution that will benefit the customer experience and also add value to companies that implement the standard. Secondly, we call upon industry and research centres, such as the Joint Research Centres (JRCs) supported by the European Commission, that have experience implementing mechanised data governance to participate in implementing and validating the proposed policy languages for terms-of-use exchanges. In particular, we seek to reduce the friction in the adoption of this Web standard by having the formal policy descriptions easily map to internal architectures for enterprise data governance.

9. Conclusion and Future Work

In this paper, we presented a comprehensive vision for the future of automated and transparent data governance on the Web, specifically focusing on the management of cookies. Our proposed solution leverages policy languages such as ODRL, Data Terms of Use (DToU), and Data Privacy Vocabulary (DPV) to describe cookie policies and data usage agreements in a machine-readable format. This approach aims to enhance user control over their privacy, streamline compliance for businesses, and facilitate regulatory oversight.

The immediate next steps involve completing the evaluation outlined in Section 7 to validate the capability of existing technologies in expressing cookie policies. This will involve detailed testing and refinement of our methodology to ensure robustness and accuracy.

In parallel, we plan to engage with legal, policy, and industry experts, as discussed in Section 8, to test the real-world viability of our proposed solution. Collaboration with regulatory bodies such as the European Data Protection Board (EDPB) and the Information Commissioner’s Office (ICO) will be crucial in ensuring that our framework aligns with regulatory requirements and can be effectively enforced.

Ultimately, such collaboration and reformation will establish a standardized (semi-)automated approach to data governance at Web scale that benefits all stakeholders – users, businesses, and regulators – and paves the way for more transparent and efficient management of personal data on the Web.

References

- [1] A. Bouhoula, K. Kubicek, A. Zac, C. Cotrini, D. Basin, Automated large-scale analysis of cookie notice compliance, 2024. URL: <https://www.usenix.org/conference/usenixsecurity24/presentation/bouhoula>, preprint.
- [2] J. A. Obar, A. Oeldorf-Hirsch, The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services, *Information, Communication & Society* 23 (2020) 128–147.
- [3] R. Iannella, S. Michael, S. Myles, V. Rodríguez-Doncel, Odrl vocabulary and expression 2.2, 2018. URL: <https://www.w3.org/TR/2018/REC-odrl-vocab-20180215/>.
- [4] R. Zhao, J. Zhao, Perennial semantic data terms of use for decentralized web (2024).
- [5] B. Esteves, D. Golpayegani, G. P. Krog, H. J. Pandit, J. Flake, P. Ryan, Data Privacy Vocabulary (DPV) v2.0, Draft Community Group Report, W3C, 2024. <https://w3c.github.io/dpv/2.0/dpv/>.
- [6] 2018 reform of eu data protection rules, 2018. URL: https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf.
- [7] eprivacy directive, https://edps.europa.eu/data-protection/our-work/subjects/eprivacy-directive_en, 2022.
- [8] S. of California Department of Justice Office of the Attorney General, California Consumer Privacy Act (CCPA) — oag.ca.gov, <https://oag.ca.gov/privacy/ccpa>, 2024. [Accessed 06-07-2024].
- [9] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web, *Scientific american* 284 (2001) 34–43.
- [10] A. V. Sambra, E. Mansour, S. Hawke, M. Zereba, N. Greco, A. Ghanem, D. Zagidulin, A. Aboulmaga, T. Berners-Lee, Solid: a platform for decentralized social applications based on linked data, MIT CSAIL & Qatar Computing Research Institute, Tech. Rep. (2016).
- [11] S. Capadisli, T. Berners-Lee, R. Verborgh, K. Kjernsmo, Solid protocol, URL <https://solid-project.org/TR/2021/protocol-20211217> (2021).
- [12] R. Verborgh, Re-decentralizing the Web, For Good This Time, 1 ed., Association for Computing Machinery, New York, NY, USA, 2023, p. 215–230. URL: <https://doi.org/10.1145/3591366.3591385>.
- [13] R. Iannella, S. Villata, Odrl information model 2.2, 2023. URL: <https://www.w3.org/TR/2018/REC-odrl-model-20180215/>.
- [14] M. G. Kebede, G. Sileno, T. Van Engers, A critical reflection on odrl, in: V. Rodríguez-Doncel, M. Palmirani, M. Araszkiewicz, P. Casanovas, U. Pagallo, G. Sartor (Eds.), *AI Approaches to the Complexity of Legal Systems XI-XII*, Springer International Publishing, Cham, 2021, pp. 48–61.
- [15] I. Akaichi, S. Kirrane, W. Slabbinck, P. Colpaert, R. Verborgh, Interoperable and continuous usage control enforcement in dataspace, 2024. URL: <https://dbis.rwth-aachen.de/SDS24/>, the Second International Workshop on Semantics in Dataspace in conjunction with the Extended Semantic Web Conference (ESWC 2024), SDS ; Conference date: 26-05-2024 Through 26-05-2024.
- [16] N. Fornara, V. Rodríguez-Doncel, B. Esteves, S. Steyskal, B. W. Smith, Odrl formal semantics, 2024. URL: <https://w3c.github.io/odrl/formal-semantics/>.

- [17] B. Esteves, H. J. Pandit, V. Rodríguez-Doncel, Odr1 profile for expressing consent through granular access control policies in solid, in: 2021 IEEE European Symposium on Security and Privacy Workshops (EuroSPW), 2021, pp. 298–306. doi:10.1109/EuroSPW54576.2021.00038.
- [18] H. J. Pandit, B. Esteves, Enhancing data use ontology (duo) for health-data sharing by extending it with odr1 and dpv, Semantic Web Preprint (2024) 1–26. URL: <https://doi.org/10.3233/SW-243583>. doi:10.3233/SW-243583, preprint.
- [19] R. Zhao, J. Zhao, Perennial semantic data terms of use for decentralized web, in: Proceedings of the ACM on Web Conference 2024, WWW '24, ACM, 2024. URL: <http://dx.doi.org/10.1145/3589334.3645631>. doi:10.1145/3589334.3645631.
- [20] H. J. Pandit, B. Esteves, G. P. Krog, P. Ryan, D. Golpayegani, J. Flake, Data privacy vocabulary (DPV) - version 2, 2024.
- [21] H. J. Pandit, Guides for Data Privacy Vocabulary (DPV), Final Community Group Report, W3C, 2024. <https://w3c.github.io/dpv/guides/>.
- [22] B. Adida, M. Birbeck, S. McCarron, I. Herman, Rdf core 1.1, 2007.
- [23] G. Kampanos, S. F. Shahandashti, Accept all: The landscape of cookie banners in greece and the uk, in: A. Jøsang, L. Fitcher, J. Hagen (Eds.), ICT Systems Security and Privacy Protection, Springer International Publishing, Cham, 2021, pp. 213–227.
- [24] M. Nouwens, I. Liccardi, M. Veale, D. Karger, L. Kagal, 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, 2020, pp. 1–13.
- [25] J. Reagle, L. F. Cranor, The platform for privacy preferences, Communications of the ACM 42 (1999) 48–55.
- [26] J. EPIC, Pretty Poor Privacy: An Assessment of P3P and Internet Privacy, <https://archive.epic.org/reports/prettypoorprivacy.html>, 2000. [Accessed 06-07-2024].
- [27] S. Zimmeck, P. Snyder, J. Brookman, A. Zucker-Scharff, Global Privacy Control — Take Control Of Your Privacy — globalprivacycontrol.org, <https://globalprivacycontrol.org/>, 2024. [Accessed 06-07-2024].
- [28] S. Zimmeck, P. Snyder, J. Brookman, A. Zucker-Scharff, Global Privacy Control (GPC), Proposal, W3C, 2024. <https://privacycg.github.io/gpc-spec/>.
- [29] R. T. Fielding, D. Singer, Tracking Preference Expression (DNT), W3C Working Group Note, W3C, 2019. <https://www.w3.org/TR/tracking-dnt/>.
- [30] S. L. Pardau, The california consumer privacy act: Towards a european-style privacy regime in the united states, J. Tech. L. & Pol'y 23 (2018) 68.
- [31] P. Voigt, A. Von dem Bussche, The eu general data protection regulation (gdpr), A Practical Guide, 1st Ed., Cham: Springer International Publishing 10 (2017) 10–5555.
- [32] S. Zimmeck, O. Wang, K. Alicki, J. Wang, S. Eng, Usability and enforceability of global privacy control, Proceedings on Privacy Enhancing Technologies 2023 (2023).
- [33] G. Bushati, S. C. Rasmusen, A. Kurteva, A. Vats, P. Nako, A. Fensel, What is in your cookie box? explaining ingredients of web cookies with knowledge graphs, Semantic Web (2023) 1–17.
- [34] D. Bollinger, K. Kubicek, C. Cotrini, D. Basin, Automating cookie consent and {GDPR} violation detection, in: 31st USENIX Security Symposium (USENIX Security 22), 2022, pp. 2893–2910.

- [35] B. Adida, et al., RDFa in XHTML: Syntax and Processing, W3C Rec. 14 Oct. 2008, W3C Semantic Web Deployment WG, 2008.
- [36] M. Florea, B. Esteves, Is Automated Consent in Solid GDPR-Compliant? An Approach for Obtaining Valid Consent with the Solid Protocol, Information 14 (2023). doi:10.3390/info14120631.
- [37] S. Luke, L. Spector, D. Rager, J. Hendler, Ontology-based web agents, in: Proceedings of the first international conference on Autonomous agents, 1997, pp. 59–66.
- [38] S. Poslad, Specifying protocols for multi-agent systems interaction, ACM Transactions on Autonomous and Adaptive Systems (TAAS) 2 (2007) 15–es.
- [39] L. F. Cranor, Designing useful and usable privacy interfaces, Talk as part of the CrySP Speaker Series on Privacy, 2021. URL: <https://www.youtube.com/watch?v=XuWAYUzUw4o>.
- [40] B. Motik, P. F. Patel-Schneider, B. Parsia, C. Bock, A. Fokoue, P. Haase, R. Hoekstra, I. Horrocks, A. Ruttenberg, U. Sattler, et al., Owl 2 web ontology language: Structural specification and functional-style syntax, W3C recommendation 27 (2009) 159.
- [41] P. Kolari, L. Ding, L. Kagal, S. Ganjugunte, A. Joshi, T. Finin, et al., Enhancing p3p framework through policies and trust, UMBC Technical Report, TR-CS-04-13 (2004).
- [42] L. Kagal, T. Berners-Lee, Rein: Where policies meet rules in the semantic web, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 2139 (2005).
- [43] S. Harris, A. Seaborne, E. Prud'hommeaux, SPARQL 1.1 Query Language, W3C Recommendation, W3C, 2013. <https://www.w3.org/TR/sparql11-query/>.