A DATASET OF GDPR COMPLIANT NER FOR PRIVACY POLICIES

Harshil Darji*, Stefan Becher, Jelena Mitrovic,
Armin Gerl, Michael Granitzer
University of Passau, Passau, Germany

Abstract

Privacy policies play a vital role in informing users about the data practices of online platforms. They are intended to help them make informed decisions regarding the processing of their personal information. Still, privacy policies are often long and complicated, making it difficult for users to understand how their data is being handled. Natural Language Processing (NLP) techniques, such as Named Entity Recognition (NER), can be employed to automatically extract meaningful information from privacy policies to ease the making of informed decisions. In this work, we present a dataset of privacy policies improved with NER annotations. The dataset consists of privacy policies from 44 online platforms. These policies were annotated to comply with the GDPR guidelines. The privacy policies are manually annotated with NER tags, highlighting relevant entities of GDPR privacy policies such as data controllers, data sources, authority, etc. We also provide the annotation guidelines used by the annotators. This annotated dataset is a valuable resource for training and evaluating NER models in the context of privacy policies.

INTRODUCTION

In our digital age, data privacy has become a crucial issue due to the widespread use of online platforms [1]. To safeguard individual rights and ensure transparent data handling, regulatory frameworks such as the General Data Protection Regulation (GDPR)[1] have been implemented. An important GDPR compliance requirement is that organizations must provide concise, transparent, intelligible, and easily accessible privacy policies, using clear and plain language to inform users about the processing of their personal information. However, in reality, privacy policies are often extensive, complicated, and hard to understand [2], making it challenging for users to comprehend the data processing procedures. Thus, a gap between the regulatory requirements and the real-life implementation of privacy policies exists due to the necessity of presenting various and extensive information on the processing of personal data, which is clearly defined, while the communication and transparency requirements are only conceptually defined and, therefore, harder to implement. Therefore, there is a need for resources that can help bridge this gap and facilitate the understanding of privacy policies for non-expert readers.

The understanding of privacy policies is crucial to protecting personal information. Natural Language Processing techniques, especially Named Entity Recognition (NER), are instrumental in identifying entities within the text, such as *Data subjects* and *Personal Data entities* [3]. However, NER has limitations in revealing complex document relationships and structures, which are essential for a thorough comprehension of privacy policies.

The GDPR policies on the web require in-depth statistical analysis. This evaluation helps users identify trustworthy policies and express their preferences. One way to improve this assessment may be to incorporate Relationship Extraction (RE). It has the potential to provide an in-depth analysis of the links between recognized entities, which can fill in any gaps left by NER. This approach can offer users a complete perspective of privacy policies.

This paper aims to address this gap in research by introducing a GDPR-compliant privacy policy dataset that has been annotated with NER tags. The dataset comprises European privacy policies from various online platforms, annotated with NER tags to identify and highlight important entities within the policies, such as Data Controller, Data Processor, Data Source, etc.

The remainder of this paper is structured as follows: Section Related work studies the related research that displays the introduction and use of similar datasets. Section Dataset introduces the dataset in question and also provides some statistics related to this dataset.

RELATED WORK

The study presented in [4] aims to provide insights into the techniques used for extracting information from textual documents and their applications by conducting a systematic mapping study on the automated analysis of privacy policies. The study analyzed 39 papers out of 1097 publications, identifying the potential for extracting individual pieces of information from privacy policies. The research addresses the growing demand for automated privacy policy analysis across various stakeholders as well as the importance of understanding privacy concerns and complying with relevant data protection laws.

The research [5] proposes PrivacyGLUE, the first benchmark for measuring general language understanding in the privacy language domain, especially focusing on privacy policies. According to this study, privacy policies need a separate benchmark due to their distinct language. PrivacyGLUE comprises seven tasks related to privacy policies and evaluates the performances of five transformer language models.

* Harshil.Darji@uni-passau.de
[1] https://gdpr-info.eu/

In the domain of data privacy, numerous datasets related to privacy policies have surfaced, requiring further studies. [6] created the OPP-115 data set, which is a collection of 115 manually annotated privacy policies, in 2016. Due to its creation date, the privacy policies the data set is based on are not compliant with the GDPR. There have been attempts to map the OPP-115 categories to GDPR articles to modernize the OPP-115 data set [7]. While this can create GDPR-compliant labels, it does not affect the outdated privacy policies as the basis of the data set. [8] annotated 350 mobile app privacy policies with privacy practices, which form the APP-350 data set in 2019. It is used to check certain compliance issues, e.g., whether a privacy policy is present, but this is limited to the privacy policies of apps. [9] created a data set by collecting over one million privacy policies, which span more than two decades, based on more than 130000 websites. They discovered interesting changes in the policies over the years, like more self-regulation and especially the impact of the GDPR. While the publicly available corpus is a good basis for investigating long-term trends, it is missing annotation for NER. All of these data sets serve a certain purpose. But there is currently no up-to-date, i.e., GDPR-compliant data set with NER annotations for categorizing data handling practices in detail.

In the context of general legal text accessibility, [10, 11] introduced annotated German legal text corpora, addressing a scarcity similar to the one reported in GDPR-compliant privacy policies. [10] introduced two German legal text corpora, addressing the lack of annotated legal resources. The first corpus is a compilation of decisions from 131 German courts, while the second is an annotated subset tailored for machine learning applications in understanding Urteilsstil. Complementing this, [11] introduced a dataset of 2944 meticulously annotated German legal references, with 21 properties each, improving legal text analysis. Their work highlights the need for annotated datasets to enhance machine readability and user comprehension. This aligns with our efforts to improve the accessibility of privacy policies through named entity recognition (NER) annotations. It highlights a shared objective across different legal fields.

The potential for improving the accessibility and understanding of privacy policies through technology is also displayed in [12, 13]. The former focuses on a structured way to categorize and analyze web pages, including privacy policies. By effectively classifying web pages, this research aids in automatically identifying privacy policies across the internet. Such capabilities are crucial for ensuring compliance with data protection laws like the General Data Protection Regulation (GDPR), as they facilitate the automated extraction of relevant information from privacy policies, aiding both users and regulatory bodies in evaluating compliance. The latter focuses on developing the OWler web crawler, a significant step in improving web crawling efficiency by focusing on topic-based content discovery, including privacy policies. This approach simplifies the process of gathering privacy policies for further analysis.

DATASET

The enactment of the GDPR in 2018 introduced stricter requirements for data privacy within the European Union. It gives users more control over their personal data by the introduction of Data Subject Rights [14, Art. 12 - Art. 23] and forced many service providers to rethink their handling of personal user data. The changes in the data handling practices directly led to a rework of existing privacy policies, in order to comply with the legal requirements of the GDPR for transparency [14, Art. 5]. This shift in the legal landscape created a research gap for a GDPR-compliant, NER-annotated data set of privacy policies because existing data sets, which were created before the enactment of the GDPR, are not applicable to the European Union anymore. We have shown, that up-to-date data sets are either missing NER tags or have another focus but web privacy policies. Therefore, we created a GDPR-compliant NER data set of web privacy policies to fix this gap.

Our data set consists of 44 European privacy policies, which have been manually annotated by legal experts. To create GDPR-compliant annotations, we have chosen the Data Privacy Vocabulary (DPV) [15], which represents the latest efforts to build a standardized ontology for privacy terms, as a basis. The DPV consists of several hierarchies, which focus on the handling of personal data as required by the GDPR, e.g., purposes, processing, or recipients. For the creation of our label set, we have chosen the most relevant entries of the DPV. Therefore, we compared several privacy policy languages, like SPECIAL [16], LPL [17], or JACPoL [18], and privacy preference languages, like YaPPL [19] or ConTra [20], in order to find a common basis of required elements. Privacy policy languages create machine-readable privacy policies, which can be further customized by the user. Privacy preference languages allow the user to define rules regarding these customization options. When a user has presented a privacy policy, represented by a privacy policy language, the preferences add support by automatically picking customization options or giving hints about mismatches. As this concept only works, if the privacy policy is machine-readable, we envision automatically translating plain-text privacy policies into such representations to enable preference matching.

Therefore, we added the following elements (based on their DPV notation), which were most commonly used in the languages we analyzed, to the label set: Data Controller (**DC**), Data Processor (**DP**), Data Protection Officer (**DPO**), Recipient (**R**), Third Party (**TP**), Authority (**A**), Data Subject (**DS**), Data Source (**DSO**), Required Purpose (**RP**), Not-Required Purpose (**NRP**), Processing (**P**), Personal Data (**PD**), Non-Personal Data (**NPD**). In addition, we analyzed the DPV for the most relevant legal terms with regard to the GDPR. Existing data sets often lack legal annotations, so with our intention to create a GDPR-compliant data set, this was an important step to take. Based on their DPV notation, the most important legal terms, regarding GDPR are Organisational Measure (**OM**), Technical Measure (**TM**),

Legal Basis (**LB**), Consent (**CONS**), Contract (**CONT**), Legitimate Interest (**LI**), Automated Decision Making (**ADM**), Retention (**RET**), Scale EU (**SEU**), Scale Non-EU (**SNEU**), Right (**RI**), Lodge Complaint (**LC**). On top of these terms, we decided to individually add the most important Data Subject Rights as labels, because the GDPR requires them to be listed in the privacy policies. This further allows for an automated compliance check. Therefore, the final labels are Art. 15 Right to access by the data subject (**DSR15**), Art. 16 Right to rectification (**DSR16**), Art. 17 Right to erasure (**DSR17**), Art. 18 Right to restriction of processing (**DSR18**), Art. 19 Notification obligations (**DSR19**), Art. 20 Right to data portability (**DSR20**), Art. 21 Right to object (**DSR21**), Art. 22 Automated individual decision-making, including profiling (**DSR22**). This results in a total of 33 categories, which form our label set. The data set consists of 33 labels with the following distribution (see Figure 1). This figure demonstrates the overall token distribution with *I-* and *B-* annotations.

Annotation guidelines

1. **Data Controller**: The individual or organization that decides (or controls) the purpose(s) of processing personal data. (E.g., *This document states the OpenStreetMap privacy policy for services formally operated and provided by the **OpenStreetMap Foundation (OSMF)**.*)

2. **Data Processor**: A *processor* means a natural or legal person, public authority, agency, or other body that processes personal data on behalf of the controller. (E.g., *We may share your data with **analytics providers**, which helps us understand how customers are using our services.*)

3. **Data Protection Officer**: An entity within or authorized by an organization to monitor internal compliance, inform and advise on data protection obligations, and act as a contact point for data subjects and the supervisory authority. (E.g., *A copy of these can be requested from the **Data Protection Officer**.*)

4. **Recipient**: A recipient of personal data can be used to indicate any entity that receives personal data. This can be a Third Party, Processor (GDPR), or Controller. (E.g., *The data collected on the systems will be accessible by the system administrators and the appropriate OSMF **working groups**.*)

5. **Third Party**: A *third party* means a natural or legal person, public authority, agency, or body other than the data subject, controller, processor, and people who, under the direct authority of the controller or processor, are authorized to process personal data. (E.g., *Cycle and Transport Map layers available via the openstreetmap.org website operated by **Gravitystorm Limited, New Malden, United Kingdom**.*)

6. **Authority**: An authority with the power to create or enforce laws or determine their compliance. (E.g., *We may disclose your data in response to official requests (e.g., court orders, subpoenas, search warrants, national security requests, etc.) ("requests") that we receive from **government authorities** or **parties to legal proceedings**.*)

7. **Data Subject**: The term *data subject* is specific to the GDPR but is functionally equivalent to the term *individual* and the ISO/IEC term *PII Principle*. (E.g., *This document is mainly intended for **OpenStreetMap contributors**.*)

8. **Data Source**: *Source* is the direct point of data collection; *origin* would indicate the original/other points where the data originates from. (E.g., *User to user messages are visible to the **sender** and recipient.*)

9. **Required Purpose**: The purpose of processing personal data required for service provision. (E.g., *We also use cookies and similar technologies **to recognize and improve your use of our websites**.*)

10. **Not-Required Purpose**: The purpose of processing personal data is not required for service provision.

11. **Processing**: The processing performed on personal data. (E.g., *When you visit this website or other websites, your **browser transmits data to our server**.*)

12. **Non-Personal Data**: The term Non-Personal Data is provided to distinguish between Personal Data and other data, indicating which data is regulated by privacy laws. (E.g., *We collect **information about your browser or application and your interaction with our website**, including (a) IP address, (b) **browser and device type**, (c) **operating system**, (d) **referring web page**, (e) **the date and time of page visits**, and (f) **the pages accessed on our websites**.*)

13. **Personal Data**: This definition of personal data encompasses the concepts used in GDPR Art.4-1 for *personal data* and ISO/IEC 27001 for *personally identifiable information (PII)*. (E.g., *The **full personal name** and **residential address** of members of the organisation.*)

14. **Organisational Measure**: Organisational measures may consist of internal policies, organizational methods or standards, and controls and audits that controllers and processors can apply to ensure the security of personal data. (E.g., *In this case, a so-called **opt-out** cookie is stored in your browser.*)

15. **Technical Measure**: Technical measures can be defined as the measures and controls afforded to systems and any technological aspect of an organization, such as devices, networks, and hardware. (E.g., *In order to protect the security of your data during transmission, we use appropriate **encryption methods in line***

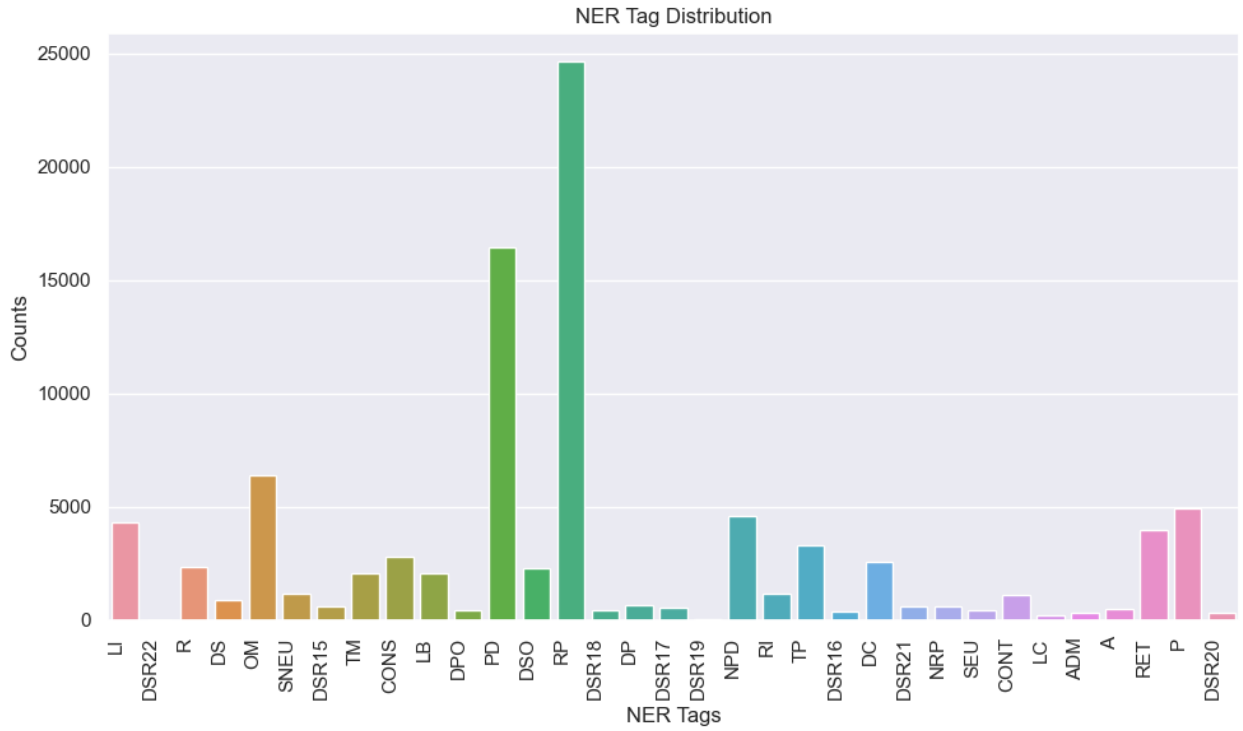

Figure 1: The number of occurrences of each NER tag in the annotated data set.

with the latest technology (e.g., SSL/TLS) *and secure technical systems.*)

16. **Legal Basis**: Legal basis (*plural: legal bases*) are defined by legislations and regulations, whose applicability is usually restricted to specific jurisdictions. (E.g., *The processing of this data is necessary for compliance with a* ***legal obligation (see GDPR article 6.1c)****.*)

17. **Consent**: Consent of the Data Subject for specified processing. (E.g., ***You can stop this behaviour*** *by explicitly turning Gravatar support off in your account settings.*)

18. **Contract**: Creation, completion, fulfillment, or performance of a contract involving specified processing. (E.g., *To our operations and working group personnel that have signed* ***confidentiality agreements****.*)

19. **Legitimate Interest**: Legitimate interests of a Party as justification for specified processing. (E.g., *We value your privacy and strive to achieve a balance between the* ***legitimate interests of the OpenStreetMap project*** *and your interests and rights.*)

20. **Automated Decision Making**: Processing that involves automated decision making. (E.g., *If you have consented to data processing or if a contract for data processing exists and* ***data processing is carried out using automated processes****.*)

21. **Retention**: Duration, temporal limitation, or condition on storage of personal data. (E.g., *Payment details for both classes of members is retained for accounting purposes* ***as long as required by law****.*)

22. **Scale EU**: Geographic coverage of processing within the European Union. (E.g., *This Section 14.2 applies only to natural persons* ***residing in the European Economic Area*** *and the United Kingdom.*)

23. **Scale Non-EU**: Geographic coverage of processing outside the European Union. (E.g., *Map tiles are provided by* ***a global network*** *of cache servers.*)

24. **Right**: The right(s) applicable, provided, or expected. (E.g., *We value your privacy and strive to achieve a balance between the legitimate interests of the OpenStreetMap project and* ***your interests and rights****.*)

25. **Lodge Complaint**: A data subject can complain to a supervisory authority if the data subject considers that the processing of personal data infringes GDPR. (E.g., *You also have the right to complain to the Bavarian state commissioner for data protection.*)

26. **Data Subject Rights** (26-33):

    - Art. 15 Right of access by the data subject
    - Art. 16 Right to rectification
    - Art. 17 Right to erasure ('right to be forgotten')
    - Art. 18 Right to restriction of processing

- Art. 19 Notification obligation regarding rectification or erasure of personal data or restriction of processing
- Art. 20 Right to data portability
- Art. 21 Right to object
- Art. 22 Automated individual decision-making, including profiling

These *Data Subject Rights* are outlined in Chapter 3 of the GDPR[2]. (E.g., *If incorrect personal data are processed, you have the* ***right to correct them (Art. 16 GDPR)***.)

Table 1 shows the entity frequency table for individual tokens. This table simply states the overall frequency of tokens available in the dataset.

Label	Frequency	Percentage
PD	4200	23.24%
P	2909	16.09%
RP	1745	9.65%
DC	1559	8.62%
NPD	955	5.28%
TP	942	5.21%
CONS	686	3.79%
TM	648	3.58%
R	585	3.24%
DS	510	2.82%
LB	419	2.32%
DSO	408	2.26%
OM	386	2.14%
LI	306	1.69%
RET	291	1.61%
SNEU	246	1.36%
RI	221	1.22%
DP	143	0.79%
CONT	129	0.71%
A	124	0.69%
ADM	109	0.60%
SEU	100	0.55%
DSR17	84	0.46%
DSR15	67	0.37%
DPO	58	0.32%
DSR16	57	0.32%
DSR21	50	0.28%
NRP	38	0.21%
DSR18	37	0.20%
LC	29	0.16%
DSR20	29	0.16%
DSR19	4	0.02%
DSR22	2	0.01%
Overall	**18076**	**100.00%**

Table 1: Entity frequency table with percentages (*rounded to two decimal places*) and overall total.

[2] https://gdpr.eu/tag/chapter-3/

The privacy policies have been reviewed by two legal experts and annotated. While annotating privacy policies, the annotators ensured proper formatting, such as line and word breaks. For inter-annotator agreement, the F1-measure between the two annotators, based on a set of 20 documents, is **0.6563** while Cohen's Kappa score is **0.6412**. Although the F1-score of **0.6563** indicates moderate agreement between annotators, it does not account for chance agreement. Cohen's Kappa, however, factors this in by underscoring the potential existence of systematic bias or inconsistencies in annotation.

The lower score is primarily the result of discrepancies in the use of Word's comment feature rather than disagreements in labeling. The decision to utilize Word's comment feature for annotating sentences or words was influenced by the annotators' familiarity with this method. When annotators highlight text for annotation, slight inconsistencies in selecting text (*including an extra space before or after a word*) can lead to discrepancies in the annotated data. These minor differences, while seemingly trivial, can affect automated processing. This affects the inter-annotator agreement scores, as it may appear that annotators disagree on the annotation of the same text when, in fact, they are aligned in their understanding but differ in their selection.

After the final annotation task, we performed a basic error analysis using Precision, Recall, and F1 scores. The results showed a precision of **0.70**, a recall of **0.62**, and an F1 score of **0.65**. To encourage further academic and practical explorations in privacy policy analysis and NER applications, our dataset is publicly accessible at the following link[3]. The dataset follows the CoNLL-2002 [21] format.

CONCLUSION

In this study, we present a dataset enriched with Named Entity Recognition (NER) annotations that comply with GDPR. It is designed to enhance the readability and accessibility of privacy policies from 44 online platforms. The Cohen's Kappa of **0.64** reflects the reliability and consistency of the annotation process but may be influenced by sentence segmentation variations. This dataset is a fundamental resource for the ongoing discussion on online privacy. Online data privacy presents dynamic challenges that require scrutiny, enhancements, and expansions.

This dataset lays the groundwork for future research in making privacy policies more accessible. By identifying key entities, subsequent research can focus on summarizing these policies, generating user-friendly interpretations, or creating visualization tools that simplify understanding privacy policies. Integrating Relationship Extraction (RE) could expand the dataset by capturing intricate relationships between entities and providing a more holistic understanding of privacy policies. We envision this corpus as a stepping stone towards these goals.

[3] https://huggingface.co/datasets/PaDaS-Lab/gdpr-compliant-ner

ACKNOWLEDGMENTS

SPONSORED BY THE

The project on which this report is based was funded by the German Federal Ministry of Education and Research (BMBF) under the funding code 01|S20049, by the project DEEP WRITE (Grant No. 16DHBKI059) and the OpenWebSearch.eu project, funded by the EU under the GA 101070014.

REFERENCES

[1] M. Smith, C. Szongott, B. Henne, and G. von Voigt, "Big data privacy issues in public social media," in *2012 6th IEEE International Conference on Digital Ecosystems and Technologies (DEST)*, 2012, pp. 1–6. 10.1109/DEST.2012.6227909

[2] D. Ibdah, N. Lachtar, S. M. Raparthi, and A. Bacha, ""why should i read the privacy policy, i just need the service": A study on attitudes and perceptions toward privacy policies," *IEEE Access*, vol. 9, pp. 166 465–166 487, 2021. 10.1109/ACCESS.2021.3130086

[3] G. B. Herwanto, G. Quirchmayr, and A. M. Tjoa, "A named entity recognition based approach for privacy requirements engineering," in *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, 2021, pp. 406–411. 10.1109/REW53955.2021.00072

[4] J. M. Del Alamo, D. S. Guaman, B. García, and A. Diez, "A systematic mapping study on automated analysis of privacy policies," *Computing*, vol. 104, no. 9, pp. 2053–2076, 2022.

[5] A. Shankar, A. Waldis, C. Bless, M. Andueza Rodriguez, and L. Mazzola, "Privacyglue: A benchmark dataset for general language understanding in privacy policies," *Applied Sciences*, vol. 13, no. 6, p. 3701, 2023.

[6] S. Wilson *et al.*, "The creation and analysis of a website privacy policy corpus," 2016, pp. 1330–1340. 10.18653/v1/P16-1126

[7] E. Poplavska, T. Norton, S. Wilson, and N. Sadeh, "From prescription to description: Mapping the gdpr to a privacy policy corpus annotation scheme," in 2020. 10.3233/FAIA200874

[8] S. Zimmeck *et al.*, "Maps: Scaling privacy compliance analysis to a million apps," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, pp. 66–86, 2019. 10.2478/popets-2019-0037

[9] R. Amos, G. Acar, E. Lucherini, M. Kshirsagar, A. Narayanan, and J. R. Mayer, "Privacy policies over time: Curation and analysis of a million-document dataset," *CoRR*, vol. abs/2008.09159, 2020. https://arxiv.org/abs/2008.09159

[10] S. Urchs., J. Mitrović., and M. Granitzer., "Design and implementation of german legal decision corpora," in *Proceedings of the 13th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART*, INSTICC, 2021, pp. 515–521. 10.5220/0010187305150521

[11] H. Darji, J. Mitrović, and M. Granitzer, "A dataset of german legal reference annotations," in *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, 2023, pp. 392–396. 10.1145/3594536.3595173

[12] M. Al-Maamari, M. Istaiti, S. Zerhoudi, M. Dinzinger, M. Granitzer, and J. Mitrovic, "A comprehensive dataset for webpage classification,"

[13] M. Dinzinger, S. Zerhoudi, M. Al-Maamari, M. Istaiti, J. Mitrović, and M. Granitzer, "Owler: Preliminary results for building a collaborative open web crawler,"

[14] E. Commission, *Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation)*, accessed: 2022-03-22, European Commission, 25, 2018. https://eur-lex.europa.eu/eli/reg/2016/679/oj

[15] H. J. Pandit *et al.*, "Creating a vocabulary for data privacy," in *On the Move to Meaningful Internet Systems: OTM 2019 Conferences*, 2019, pp. 714–730.

[16] S. H2020, *Scalable policy-aware linked data architecture for privacy, transparency and compliance*, accessed: 2022-03-22, SPECIAL H2020. https://www.specialprivacy.eu/

[17] A. Gerl, N. Bennani, H. Kosch, and L. Brunie, "Lpl, towards a gdpr-compliant privacy language: Formal definition and usage," in *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXXVII*, 2018, pp. 41–80.

[18] H. Jiang and A. Bouabdallah, "Jacpol: A simple but expressive json-based access control policy language," in *Information Security Theory and Practice*, 2018, pp. 56–72.

[19] M.-R. Ulbricht and F. Pallas, "Yappl - a lightweight privacy preference language for legally sufficient and automated consent provision in iot scenarios," in *DPM/CBT@ESORICS*, 2018, pp. 329–344. 10.1007/978-3-030-00305-0_23

[20] S. Becher and A. Gerl, "Contra preference language: Privacy preference unification via privacy interfaces," *Sensors*, vol. 22, no. 14, 2022. 10.3390/s22145428

[21] E. F. Tjong Kim Sang, "Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition," in *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002. https://aclanthology.org/W02-2024