From Prescription to Description: Mapping the GDPR to a Privacy Policy Corpus Annotation Scheme

Ellen POPLAVSKA [a], Thomas B. NORTON [b], Shomir WILSON [a], and Norman SADEH [c]

[a] *Pennsylvania State University, University Park, Pennsylvania, USA*
[b] *Fordham University School of Law, New York, New York, USA*
[c] *Carnegie Mellon University, Pittsburgh, Pennsylvania, USA*

Abstract. The European Union's General Data Protection Regulation (GDPR) has compelled businesses and other organizations to update their privacy policies to state specific information about their data practices. Simultaneously, researchers in natural language processing (NLP) have developed corpora and annotation schemes for extracting salient information from privacy policies, often independently of specific laws. To connect existing NLP research on privacy policies with the GDPR, we introduce a mapping from GDPR provisions to the OPP-115 annotation scheme, which serves as the basis for a growing number of projects to automatically classify privacy policy text. We show that assumptions made in the annotation scheme about the essential topics for a privacy policy reflect many of the same topics that the GDPR requires in these documents. This suggests that OPP-115 continues to be representative of the anatomy of a legally compliant privacy policy, and that the legal assumptions behind it represent the elements of data processing that ought to be disclosed within a policy for transparency. The correspondences we show between OPP-115 and the GDPR suggest the feasibility of bridging existing computational and legal research on privacy policies, benefiting both areas.

Keywords. privacy, privacy laws, privacy policies, theory, annotation, GDPR, General Data Protection Regulation

Introduction

In 2018, the GDPR entered into force, becoming one of the most influential privacy laws to date. As a result, businesses and organizations were required to change their privacy protocols to comply. For many, these changes included changes to the privacy policies provided to users. In particular, many businesses and organizations were compelled to update their privacy policies to state specific information about their data practices.

Recent efforts in natural language processing (NLP) have addressed the demand for automatic information extraction from privacy policies to ease legal analysis and build privacy-enhancing consumer technologies [15,8,4]. This work requires the creation of privacy policy corpora that contain annotations identifying salient details about privacy practices. Currently, the most extensive text annotation scheme dedicated to privacy poli-

cies is the OPP-115 annotation scheme [14], which was initially created for a corpus of 115 annotated privacy policies. This corpus now appears in several projects as part of tasks to extract information from privacy policies [2,3,9,10]. The annotation scheme was created to be agnostic to particular laws, instead concentrating on a general concept of privacy practices, or activities that an organization may perform with customers' information. Determining the relevance of OPP-115 to the GDPR clarifies how well existing work based upon this annotation scheme addresses the concerns of modern privacy law.

We perform a comparative study of the OPP-115 annotation scheme with the GDPR Article 5 principles for processing of personal data, as well as other relevant articles of the GDPR, identifying matches and mismatches between these two systematizations. We show strong connections between the two, validating OPP-115's applicability and the relevance of NLP research that continues to use the annotation scheme. We release our dataset of connections between the GDPR and OPP-115 to promote further NLP research to automatically identify connections between privacy policies and privacy law.[1]

1. Related Work

1.1. OPP-115 and its Uses

The Online Privacy Policies, Set of 115 (OPP-115) Corpus released by Wilson et al. [14] contains 115 privacy policies annotated by law students. It provides an annotation scheme of ten mutually exclusive categories into which segments of privacy text, known as *data practices*, may be sorted. The OPP-115 corpus and its annotation scheme have been utilized by other privacy researchers. Sathyendra et al. [9] used the corpus to train models to extract opt-out choices from privacy policies. Harkous et al. [2] used the corpus to classify privacy practices and answer non-factoid questions. Story et al. [10] used the corpus to automatically identify opt-out choices on websites and locate potential noncompliance. Mousavi et al. [3] used the corpus to predict categories for paragraphs of privacy text. Researchers have continued to use this annotation scheme to represent the structure of a standard privacy policy. To date, however, there has been no published work analyzing how accurately the OPP-115 categories represent privacy legislation.

1.2. Computational Uses of the GDPR

Since the GDPR came into effect, researchers have considered methods to determine compliance. Truong et al. [13] have envisioned a personal data management platform designed around GDPR compliance. Tesfay et al. [11] have created PrivacyGuide, a tool that classifies privacy policy content into eleven aspects constructed around GDPR compliance. Torre et al. [12] have created a UML representation of the GDPR as a first step towards automated compliance checking. Palmirani et al. [5] have proposed a framework for modelling legal documents for compliance checking. Palmirani et al. [6] have developed PrOnto, a privacy ontology modelling the conceptual cores of the GDPR. Bonatti et al. [1] have created the SPECIAL Usage Policy Language to describe cores of GDPR-compliant usage policies. Polleres et al. [7] have created the Data Privacy Vocabulary to describe and categorize GDPR-compliant personal data handling. In contrast with others' work, ours fills a theoretical gap between privacy policy annotations and uses of AI and NLP on privacy policies. Additionally, the OPP-115 annotation scheme's use beyond one project motivates further examination of how it connects with specific privacy laws.

[1] `usableprivacy.org/data/`

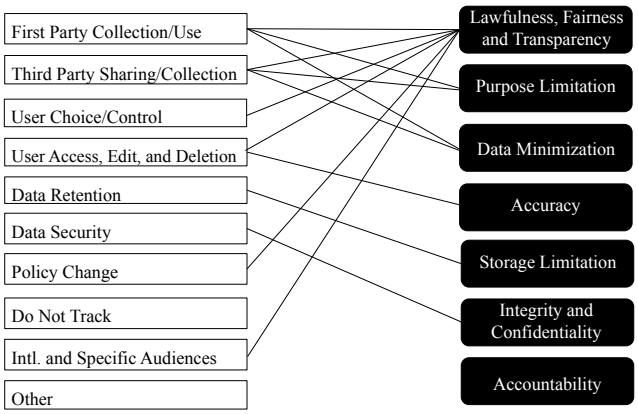

Figure 1. OPP-115 categories, left, connected to principles from GDPR Article 5, right.

2. Approach

In Article 5, the GDPR details a set of principles for data processing, which provide an overview of the regulation's expectations for data controllers and processors. We compare these principles to the categories of OPP-115, which represent the most general level of the annotation scheme, and identify thematic connections. These connections represent instances when the principles and categories codify the same expectations (prescriptive and descriptive, respectively) for the contents of privacy policies. We also create a dataset of the connections between the 99 articles of the GDPR and the categories of OPP-115. In developing these associations, we consider the definitions of each category of OPP-115, the descriptions of the articles, the audience of each particular article, and whether the concepts described in a particular article might belong in a privacy policy.

3. Results and Discussion

Of the 99 articles, we find associations with categories of OPP-115 within 49. We find a total of 88 connections between GDPR articles and OPP-115 categories. 88 of these occur within the first five chapters of the GDPR, suggesting that some chapters contain more pertinent privacy policy details than others. Most articles are associated with multiple categories. The median number of connections for an article is two, demonstrating that the concepts within each article are usually applicable to multiple categories and that GDPR concepts overlap considerably across sections. Figure 1 displays connections between OPP-115 categories and GDPR principles. These represent thematic similarities between the concepts guiding the GDPR and the categories for data practices described by OPP-115. We release the full set of connections in CSV format for further research.

These connections and gaps between the OPP-115 annotation scheme and the GDPR reflect the similarities and differences between what privacy experts believed were the essential components of privacy policies in 2016 and the codified European privacy regulation of 2018. These give insight as to how accurately OPP-115 legal scholars' observations reflect today's legislative understanding of privacy concepts. Comparing the principles of the GDPR to the categories of data practices in OPP-115, it is apparent that legal scholars' decisions about categories of data practices are similar to legislators' descrip-

tions of similar concepts. While OPP-115 separates First Party Collection/Use and Third Party Sharing/Collection, the GDPR presents principles that apply to all data processing by controllers and processors. This may reflect the fact that OPP-115 was created to sort data practices in privacy policies, where first-party and third-party processing are often listed in distinct sections, while the GDPR provides guidance for all data processing.

In addition to revealing how the legal insights behind OPP-115 reflect recent privacy regulation, this work demonstrates how accurately the OPP-115 corpus and annotation scheme currently used by researchers represent it. This allows researchers to contextualize their results within a set of principles similar to those represented in the regulation.

4. Acknowledgements

This work was supported in part by the US National Science Foundation under Grants #CNS-1914444, #CNS-1914446, and #CNS-1914486, as well as the PA Space Grant Consortium Research Internship Program.

References

[1] P. Bonatti, S. Kirrane, I. M. Petrova, L. Sauro, and E. Schlehahn. *The SPECIAL Usage Policy Language*, 2019. `https://ai.wu.ac.at/policies/policylanguage/`.

[2] 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 *Proc. USENIX Security*, 2018.

[3] N. Mousavi, D. Graux, and D. Collarana. Towards measuring risk factors in privacy policies. In *AIAS@ICAIL*, 2019.

[4] A. Oltramari, D. Piraviperumal, F. Schaub, S. Wilson, S. Cherivirala, T. B. Norton, N. C. Russell, P. Story, J. Reidenberg, and N. Sadeh. Privonto: A semantic framework for the analysis of privacy policies. *Semantic Web*, 2017.

[5] M. Palmirani and G. Governatori. Modelling legal knowledge for GDPR compliance checking. In *Proc. JURIX*, 2018.

[6] M. Palmirani, M. Martoni, A. Rossi, C. Bartolini, and L. Robaldo. Legal ontology for modelling GDPR concepts and norms. In *JURIX*, 2018.

[7] A. Polleres, B. Bos, B. Bruegger, E. Kiesling, E. Schlehahn, F. Ekaputra, H. Pandit, J. Fernández, M. Lizar, and R. Hamed. *Data Privacy Vocabulary (DPV)*, 2020. `https://dpvcg.github.io/dpv/`.

[8] A. Ravichander, A. W. Black, S. Wilson, T. B. Norton, and N. Sadeh. Question answering for privacy policies: Combining computational and legal perspectives. In *Proc. EMNLP-IJCNLP*, 2019.

[9] K. M. Sathyendra, F. Schaub, S. Wilson, and N. Sadeh. Automatic extraction of opt-out choices from privacy policies. In *Proc. AAAI Symposium on Privacy-Enhancing Technologies*, AAAI Fall Symposium - Technical Report, 2016.

[10] P. Story, S. Zimmeck, A. Ravichander, D. Smullen, Z. Wang, J. Reidenberg, N. C. Russell, and N. Sadeh. Natural language processing for mobile app privacy compliance. In *Proc. PAL*. CEUR Workshop Proceedings, 2019.

[11] W. B. Tesfay, P. Hofmann, T. Nakamura, S. Kiyomoto, and J. Serna. PrivacyGuide: Towards an implementation of the EU GDPR on internet privacy policy evaluation. In *Proc. IWSPA*, 2018.

[12] D. Torre, G. Soltana, M. Sabetzadeh, L. Briand, Y. Auffinger, and P. Goes. Using models to enable compliance checking against the GDPR: An experience report. In *Proc. MODELS*, 2019.

[13] N. B. Truong, K. Sun, G. M. Lee, and Y. Guo. GDPR-compliant personal data management: A blockchain-based solution. *IEEE Transactions on Information Forensics and Security*, 2020.

[14] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. Giovanni Leon, M. Schaarup Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell, T. B. Norton, E. Hovy, J. Reidenberg, and N. Sadeh. The creation and analysis of a website privacy policy corpus. In *Proc. ACL*. Association for Computational Linguistics, 2016.

[15] S. Wilson, F. Schaub, R. Ramanath, N. Sadeh, F. Liu, N. Smith, and F. Liu. Crowdsourcing annotations for websites' privacy policies: Can it really work? In *Proc. WWW*, 2016.