From Prescription to Description: Mapping the GDPR to a Privacy Policy Corpus Annotation Scheme
Readme for Dataset Release v1.0

Usage Information

The annotations are made available for research, teaching, and scholarship purposes only, with further parameters in the spirit of a Creative Commons Attribution-NonCommercial License.

If you use this dataset for research, you should cite the following paper:

Ellen Poplavska, Thomas B. Norton, Shomir Wilson, and Norman Sadeh. From Prescription to Description: Mapping the GDPR to a Privacy Policy Corpus Annotation Scheme. In Proceedings of the 33rd International Conference on Legal Knowledge and Information Systems (JURIX), December 9-11, 2020.

The above paper is also an essential read for understanding the contents of the dataset.

Dataset Contents

This dataset contains comparisons between the annotation scheme of a corpus of annotated privacy policies known as Online Privacy Policies, Set of 115 (OPP-115) and two different levels of the European Union's General Data Protection Regulation (GDPR). These two levels are the principles related to processing of personal data detailed in seven subsections of Article 5 of the GDPR, as well as the entire set of 99 Articles contained within the GDPR. The annotation scheme of OPP-115 is used here as it has been utilized by privacy researchers to represent the components of a typical complete privacy policy. This annotation scheme was developed by legal scholars and privacy researchers to sort segments of text within privacy policies, known as data practices, into one of ten mutually exclusive categories. The GDPR is used as it is one of the most extensive and influential pieces of data privacy legislation to date, with a scope impacting millions of businesses and organizations across the globe. Because the annotation scheme of OPP-115 was published before the GDPR was passed, it represents the legal intuition of privacy scholars regarding the essential components of privacy policies before the influence of this piece of legislation, and may be compared to the principles and Articles that legislators regarded as essential and codified in the GDPR.

This dataset consists of three files. The first, *categories_principles_matrix.csv*, is a set of connections between the categories of OPP-115 and the principles of the GDPR. The second, *categories_articles_matrix.csv*, is a set of connections between the categories of OPP-115 and the Articles of the GDPR. These comparisons demonstrate the thematic correspondences between the principles guiding the entirety of the GDPR, as well as the individual Articles of the GDPR, and the expected contents of a privacy policy as represented by the OPP-115 annotation scheme. The third file, *connections_overview.csv*, contains an alternate representation of the same information in the first two sets of connections. This third file may be easier for a human reader to understand, while the first two files are intended for programmatic parsing.

For more information regarding the OPP-115 annotation scheme, please refer to the following paper:

The creation and analysis of a website privacy policy corpus. Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, August 2016.

To read the GDPR, please refer to the document webpage maintained by the European Union: https://gdpr-info.eu/

Credits

The annotations were created by Ellen Poplavska, with input from the co-authors. This readme was written by Ellen Poplavska and Shomir Wilson.

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.