Creation and Analysis of an International Corpus of Privacy Laws

Sonu Gupta[†*] Geetika Gopi[‡*] Harish Balaji[‡*] Ellen Poplavska[†]
Nora O'Toole[†] Siddhant Arora[‡] Thomas Norton[¶]
Norman Sadeh[‡] Shomir Wilson[†]
[†]Pennsylvania State University [‡]Carnegie Mellon University [¶]Fordham University
State College, PA; Pittsburgh, PA; Bronx, NY
tnorton1@law.fordham.edu, nsadeh@cs.cmu.edu, shomir@psu.edu

Abstract

The landscape of privacy laws and regulations around the world is complex and ever-changing. National and super-national laws, agreements, decrees, and other government-issued rules form a patchwork that companies must follow to operate internationally. To examine the status and evolution of this patchwork, we introduce the Privacy Law Corpus, of 1,043 privacy laws, regulations, and guidelines, covering 183 jurisdictions. This corpus enables a large-scale quantitative and qualitative examination of legal focus on privacy. We examine the temporal distribution of when privacy laws were created and illustrate the dramatic increase in privacy legislation over the past 50 years, although a finer-grained examination reveals that the rate of increase varies depending on the personal data types that privacy laws address. Our exploration also demonstrates that most privacy laws respectively address relatively few personal data types. Additionally, topic modeling results show the prevalence of common themes in privacy laws, such as finance, healthcare, and telecommunications. Finally, we release the corpus to the research community to promote further study.

Keywords: Privacy Law, Corpus, Personally identifiable information, language resource creation, text analysis

1. Introduction

Privacy is a growing topic of attention for legislative and regulatory bodies around the world, and a growing number of documents produced by governments provide instructions for this topic. These government-issued instructions include legally binding documents such as laws and regulations, and non-legally binding documents such as guidelines for following laws. Legal jurisdictions around the world have their own sets of privacy laws, shaping the legal framework surrounding privacy within their particular jurisdictions.

At the same time, text analysis techniques have made it possible to study legal texts on a large scale. Prior efforts have studied legal text about privacy in the form of privacy policies, yielding insights for legal scholars and language models for the creation of privacy-enhancing technologies (Hosseini et al., 2021; Ravichander et al., 2021; Wilson et al., 2018). Other efforts have applied NLP to legal text in the context of knowledge graphs and text interpretation (Moreno-Schneider et al., 2020; Robaldo et al., 2019). However, despite the growing interest in privacy law, NLP researchers have lacked a large-scale collection of privacy documents from around the world. This stems from the nontrivial effort necessary to produce such a collection. Often there are several official and unofficial versions of a privacy document on the web. Governments often publish instructions or guideline documents pertaining to these laws[1] [2] which makes it challenging to distinguish them from legally enforceable documents. The task is further exacerbated by the absence of official translations of these laws.

We address these challenges and present a *Privacy Law Corpus*[3]. To the best of our knowledge, this corpus is the most comprehensive corpus of privacy laws to date, with natural language text presented in original languages and English. The texts are also paired with document metadata such as source URLs, applicable jurisdictions, dates of enactment, relation to international agreements, and other significant information. We coin the term Privacy Law Corpus to characterize these documents, as the corpus encompasses laws, regulations, and government-issued guidelines and recommendations intended to instruct citizens, organizations, law enforcement, or lawmakers on required actions to protect digital privacy.

We also present the first large-scale study of privacy laws using text analysis techniques. We examine the temporal and topical trends in privacy

[*] The first three authors contributed equally.

[1] https://www.oaic.gov.au/__data/assets/pdf_file/0012/8013/privacy-safeguardcombined-chapters.pdf

[2] https://www.priv.gc.ca/en/privacy-topics/technology/online-privacy-trackingcookies/tracking-and-ads/gl_ba_1112/

[3] https://anonymous.4open.science/r/Privacy-Law-Corpus-CDFD/

LREC-COLING 2024, pages 4092–4105
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

laws, showing a dramatic increase in attention to privacy over the past 50 years, a varied and nuanced distribution of mentions to personal information types, and a set of common themes that privacy laws address.

2. Related Work

We describe prior efforts toward language resource creation and NLP applications on four related domains of text: laws in general, legal documents, privacy policies, and privacy laws. **Law Corpora:** Prior work has created international law corpora with varying foci. (Elliott, 2011) curated a master list of 779 international human rights instruments from 1863 to 2003 to highlight significant violations of those rights. Lame (Lame, 2005) proposed an NLP-based technique to extract concepts and relations from 57 French codes gathered from government websites that constitute 59,000 articles.

Legal Document Corpora: The analysis and interpretation of text dominates the field of law. Lawyers, judges, and regulators continuously compose legal documents such as memos, contracts, patents, and judicial decisions. Accordingly, there is a body of research about creating corpora of such legal documents. These corpora facilitate building Natural Language Understanding (NLU) technologies to assist legal practitioners. (Malik et al., 2021) introduced a corpus of Indian legal documents toward building an automated system for predicting the outcome of a legal trial as well as explaining the outcome. These automated systems can assist judges and help expedite the judicial process.

Privacy Policy Corpora: Over the last decade, there has been significant growth in research about online privacy policies (i.e., natural language statements about data practices that organizations are required to post on their websites or for their mobile apps). The existence of data and high-quality annotations are essential for the application of both natural language processing and crowd-sourcing techniques to address the challenges posed by online privacy policies. This requirement has generated two threads in online privacy policy research: (i) annotation of privacy policy documents to facilitate future analysis and (ii) large-scale collection and analysis of privacy policies. The initial annotation attempts involved manual annotation of privacy policies by legal experts and crowd workers. Two such corpora are OPP-115 (Wilson et al., 2016) and APP-350 (Zimmeck et al., 2019). Although these corpora are relatively small, their annotations enabled several researchers to use them to train machine learning models to extract salient details from privacy policies (Shvartzshanider et al., 2018).

Privacy Law Corpora: In 2011, Graham Greenleaf conducted the initial worldwide data privacy law survey, identifying 76 countries meeting minimum international data privacy standards. A decade later, the seventh edition of this survey expanded the global list to include 145 countries with Data Privacy Laws and 23 with pending bills (Greenleaf, 2021).

Similarly, in (Greenleaf, 2014), the author discussed and analyzed Asian data privacy laws in-depth. Our work closely aligns with the previous work by Greenleaf. We take a broader perspective of the data protection laws and broaden the inclusion criteria to extend our corpus by including more jurisdictions and documents (e.g., guidelines). In addition, all the above efforts present only qualitative analysis. In contrast, we employ both quantitative and qualitative methodologies. We also leverage NLP tools and machine learning algorithms to study this large-scale corpus. Lastly, unlike previous work that shared the list of the names of these documents, we share the original text of all the documents. We also consider multilinguality and share both the original non-English text and English translations when applicable.

3. Corpus Creation

Corpus creation required a series of overarching tasks: searching by jurisdiction for document that ought to be included in the corpus, determining precise jurisdiction and document inclusion criteria, manually collecting privacy laws for the selected jurisdictions from the internet, and categorizing these documents into three subdivisions. We summarize the entire pipeline of the corpus creation tasks in **Figure 1**.This is the first large-scale research to build and analyze a comprehensive privacy law corpus. Corpus creation was a challenging task and required significant legal and technical expertise due to the heterogeneous nature of privacy laws and regulatory documents.

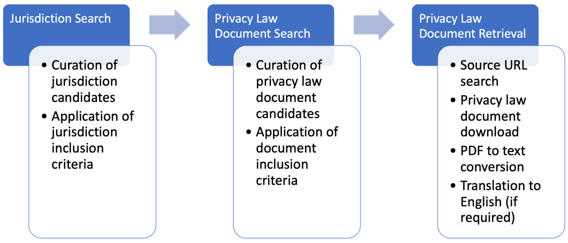
Jurisdiction Search
• Curation of jurisdiction candidates
• Application of jurisdiction inclusion criteria
Privacy Law Document Search
• Curation of privacy law document candidates
• Application of document inclusion criteria
Privacy Law Document Retrieval
• Source URL search
• Privacy law document download
• PDF to text conversion
• Translation to English (if required)

Figure 1: The end-to-end pipeline of the creation of the Privacy Law Corpus.

3.1. Jurisdiction

Intending to achieve extensive coverage of nation-level jurisdictions worldwide, we curate a list of candidate jurisdictions prior to collecting their privacy law documents. First, to build this list, we defer to the existing work by (Greenleaf, 2019), and leading legal experts such as Data Guidance (dat, 2022) and DLA Piper (dla, 2022). For the sake of simplicity, we call them *reasoning documents*. Next, we instate a series of inclusion criteria to scope our list of jurisdictions.

We include a jurisdiction if it is either a recognized member (or observer) state of the United Nations or falls under a special category. For more details about the jurisdiction inclusion criteria, *refer to Appendix A*.

In summary, the process results in a list of 183 jurisdictions, with 161 at the country level (86% of 193 United Nations member states (un, 2022)). For the remaining 27 countries in United Nations member states, either no privacy laws exist or they are irretrievable from the web.

3.2. Privacy Law Documents

To create an initial list of candidate documents, we refer to the list of 132 privacy laws collected by (Greenleaf, 2019) and documents from online sources compiled by legal experts pertaining to each applicable jurisdiction.

We developed two sets of rules for document inclusion:

Based on document type: The document is required to satisfy at least one criterion related to enforceability requirements and content. (Refer to Appendix B for more details). Based on source type: The document is required to satisfy at least one criterion related to comprehensiveness, release entity, and intended audience. (Refer to Appendix C).

For a detailed description of excluded document types, *refer to Appendix D.*

3.3. Document Collection

If a document is deemed fit to be included within the corpus, its web source is identified from references in the reasoning documents. However, in most cases, reasoning documents do not cite web sources, especially for non-English documents. In such cases, we seek to download both the original version and, if available, a human-translated English version of the document. The document collection process includes locating, downloading, and uploading documents to the repository and recording metadata. All documents were downloaded in PDF format to preserve visual formatting. Collected PDF documents were converted into text files (.txt) using Apache Tika (Team, 2021). We attempted to use OCR technology (Mori et al., 1999) to convert scanned documents to text, although the software was not always successful. In instances when we were only able to collect a non-English version of a document, we translated its text file to English using online tools. We elaborate on the issues and challenges of the translation process in section 4.1.

3.4. Subdivision of the Corpus

We divided the initial version of the Privacy Law Corpus into three sets based on the document's availability for text analysis: the primary set, the untranslatable set, and the irretrievable set. The 'primary set' comprises documents within the canonical body of the corpus. It includes documents that are available in English, regardless of whether they were originally available in English, human-translated, or machine-translated into English. Next, as the name suggests, the 'untranslatable set' includes non-English document text that couldn't be translated into the English language. Lastly, the 'irretrievable set' includes documents that we intended to include in the corpus but did not find an appropriate web source for retrieval. The final published version of the corpus includes the primary and untranslatable set of privacy law documents. The failure modes for the untranslatable set and the irretrievable set are detailed in *Appendix D*. In addition, if available, we retrieve the metadata (e.g., the year of enactment, if in effect or repealed, etc.) for all the documents in the three sets for temporal analysis purposes.

Set Name	# of Docs	Percentage
Primary Set	1,043	87.21%
Untranslatable Set	14	1.17%
Irretrievable Set	139	11.62%

Table 1: Summary of the subdivision of the documents.

4. Distribution of Privacy Laws

The process described in the above section results in 1,043 documents for analysis. Based on the coverage, we classify the jurisdictions into three categories, (1) National, (2) International, and (3) State/Province. As shown in **Table 3**, the majority of the documents cover national level jurisdictions and contribute to 95.11% of the documents in the corpus with 183 unique jurisdictions, whereas 161 distinct countries make up 90.89% of the documents. Countries that participated in various international agreements also have their own unique sets of documents within the jurisdiction of their own country.

The collected documents are laws and regulations, rules, guidelines, and other government-released documents, communiques, notices, circulars, orders, decrees, and decisions. Each document is in its current state or the latest state of revision if any. The latest revision date is recorded for each document from the law category. We add the promulgation date as the last revision date if no revisions have occurred.

We show the number of enforceable privacy laws in our corpus per country as a map in **Figure 2**. We observe that Turkey has the most privacy laws, followed by Japan, Uzbekistan, and France. We explore the reason for Turkey's exceptional number in Section 2.

4.1. Translations

The corpus comprises documents in 54 languages, with 37.12% documents in English, making it the most common language. In addition, 85 documents are written in both English and the native language of the region in a single document. Chad is the only exception where the document is written in two languages (French/Arabic) and neither of them is English. There are six languages that only appear in combination with the English language. For instance, all eleven documents that contain Maltese, also contain English.

We attempt to create English translations for all the non-English documents in the corpus, to establish a uniform natural language for text analysis. Based on the translation, we divide the corpus into four classes: (1) Originally in English, (2) Official translation, (3) Unofficial translation, and as the name suggests, and (4) Non-Government Machine. If a document is in a language other than English, we seek an official English translation provided by the source of the official non-English document. Sometimes, official sources provide a translated version but call it an unofficial document for legal purposes. In the absence of the availability of such translations, we turn to international privacy expert sites, with exact sources noted in the corpus metadata. However, if the translation is still unavailable, we use translation tools like Google Translate (phi, 2022). Machine translated documents are referred to as 'non-government machine'. As we show in **Table 2**, 53.65% of documents within the corpus are machine translated. Out of 654 non-English documents, we utilize their English titles to locate the source of the English version of the document on the web. In the absence of non-English titles, we turn to Google translate. However, it fails to provide a usable translation for a few titles in the Russian language. Therefore, we turn to Yandex Translate (yan, 2022).

Translation Type	# of Docs	(%) Total
Non-Government Machine	558	53.65
Originally English	386	37.12
Official Translation	66	6.35
Unofficial Translation	19	1.83
Originally in other languages	11	1.06

Table 2: Distribution of the sources of English translations.

4.2. Temporal Distribution

We examine the distribution over time of the creation of privacy laws, as the corpus contains documents dated as early as 1803 and as recently as December 2020. We illustrate the pace of privacy laws enacted over this date range in **Figure 3**. It is a dual-vertical axis graph where the left and right vertical axes show the cumulative and total number of privacy laws enacted over the years, respectively. We perform a chi-square test for the goodness of fit and detect a trend, starting in 1966, that the pace of privacy laws grows exponentially with rate parameter (λ) equal to 0.054. We also find a sharper exponential growth in the 21st century with (λ) equal to 0.036. It should be noted that in a few documents first written in the late 80s and early 90s, the data privacy statements were included only after revisions. For example, the criminal code of Finland was enacted in 1889, but a data privacy section was not added to it until 2015. We notice two discernible peaks in 2016 and 2018. In 2016, 86 privacy laws were issued, out of which 32 were published in Turkey only. After the failed July 15, 2016 coup attempt (tur, 2023) in Turkey, several emergency decrees were published (Malaurie et al., 2016). We speculate that the coup attempt was the cause of the sudden increase in privacy laws in Turkey. We also observe that the largest number of privacy laws were issued in 2018, with a total of 131 privacy laws in 67 distinct jurisdictions. We speculate that the enactment of GDPR in early 2018 may have caused the increase in the new documents as 16.79% of documents explicitly mention GDPR in their title. Additionally, GDPR may have encouraged the presence of documents to be in digital format and available over the web. We note a continuous increase every decade, with the largest number of jurisdictions (62) receiving their first privacy laws between 2010 and 2019.

5. Text Analysis

We employ text analysis methods to study the Privacy Law Corpus, examining its composition and trends. We interchangeably refer to it as the 'corpus' or 'primary set'. **Table 4** offers summary statistics, revealing document lengths ranging from 46 to 590,085 words.

5.1. Personally Identifiable Information

Personally identifiable information (PII) includes any information associated with an identified or identifiable living person, in particular, that can be

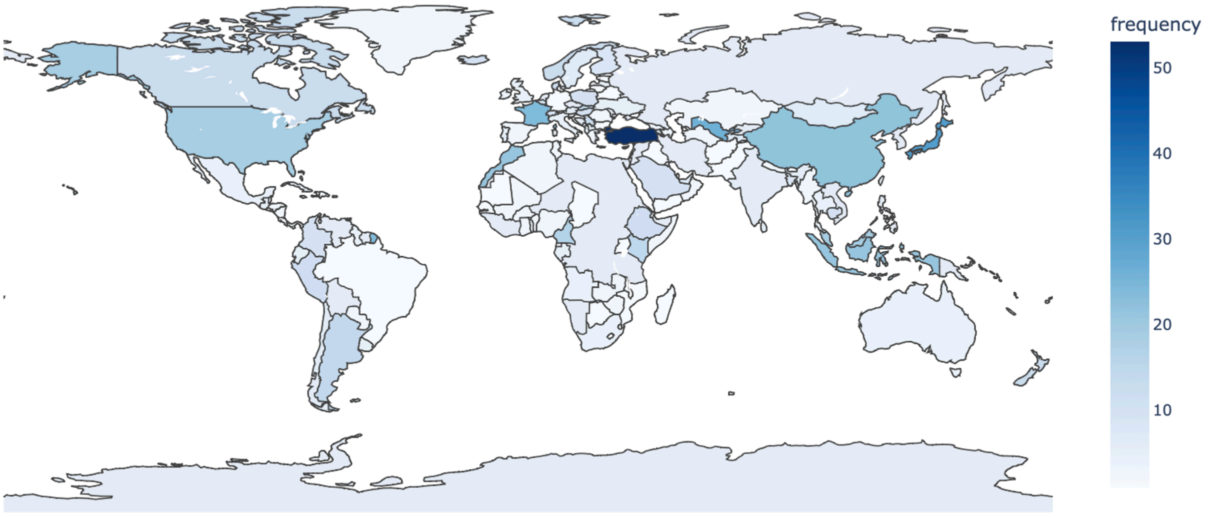

Figure 2: World map representing the number of enforceable privacy laws from each country in the Privacy Law Corpus.

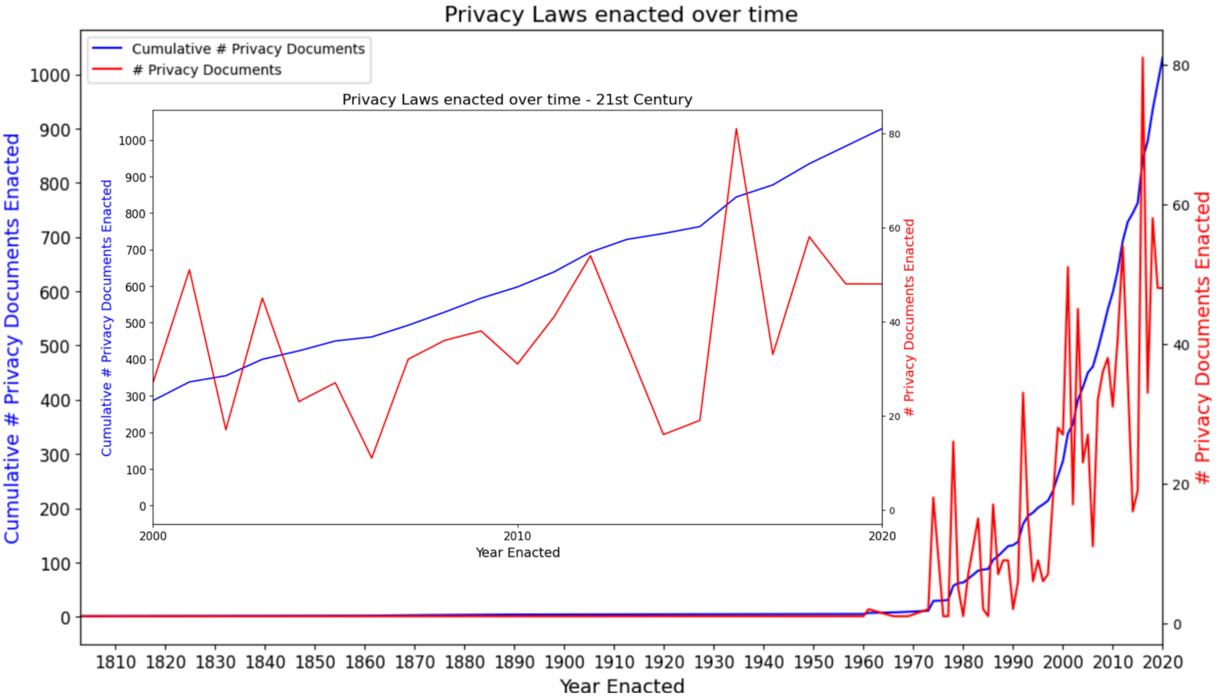

Figure 3: Privacy laws enacted over time.

connected to an identifier such as a name, national identification number, email address, and more (Commission, 2021). Privacy laws often explicitly include a descriptive definition of PII at the beginning of the document.

We examine the distribution of mentions of PII types in privacy laws and their trends over time to identify differences in attention and temporal trends. To do this, we create a list of 183 PII keywords with the help of following official sources:

- *More Data Types More Problems: A Temporal Analysis of Complexity, Stability, and Sensitivity in Privacy Policies* (Mori et al., 1999)
- The U.S. National Archives and Records Administration (Archives and Administration, 2022)
- The U.S. National Institute of Standards and Technology (NIST, 2022)
- The U.S. Federal Trade Commission (FTC)
- The European Commission (Eur, 2022)

We expand the list by including several country-

Jurisdiction Type	Coverage Type	# UJ	# Docs	Examples
Countries	N	161	948	Albania
British Overseas Territories	N	3	24	Cayman Islands
Crown Dependencies	N	3	20	Isle of Man
Special Administrative Regions	S/P	3	18	Macau
International Organizations	I	4	14	United Nations
Special Economic Zones	S/P	4	8	Qatar Financial Centre
Intergovernmental Organizations	I	4	5	US + 23 Countries
State	S/P	1	3	California(USA)

Table 3: Summary of corpus composition. In the Coverage Type column, N, I, and S/P represent National, International, and State/Province jurisdictions, respectively. #UJ is the number of unique jurisdictions.

Mean	16,399
Minimum words in a file	46
Maximum words in a file	590,085
Median	6,715
Total words	17.03M

Table 4: Summary of the subdivision of the documents.

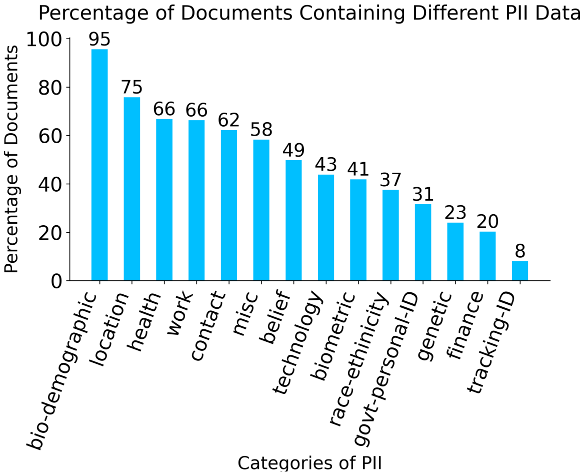

Figure 4: Distribution of the number of PII types present per document.

specific alternates for each PII keyword. For example, national ID schemes are known by many names, including Social Security Number in the US, Documento Nacional de Identidad in Argentina, and Aadhaar in India (Forum, 2021). We also account for variations in non-country-specific terms, like different name formats (e.g., middle name, first name, last name).

We classify PII keywords into 14 categories, detailed in **Appendix F**, with a miscellaneous category for unclassified terms. We standardize keywords to lowercase and singular forms (e.g., “ges” to “age”) and account for abbreviations (e.g., “mobile number” and “mobile no.”). These preprocessing steps are applied to the corpus before statistical analysis.

In **Figure 4** we observe that, apart from the biographical category, the presence of keywords from each category in documents is significantly lower. Additionally, we note that the tracking-ID category, which includes technical terms, is infrequently found in privacy laws. This suggests that laws refrain from committing to regulating specific representation of the information. About 60.79% of privacy laws represent three or fewer PII types. Only 0.19% privacy laws (2 documents) are comprehensive enough to cover the 14 PII types listed in the **Appendix F**. According to this measure, the two most comprehensive privacy laws are California Consumer Privacy Act (CCPA) and California Privacy Rights Act (CPRA).

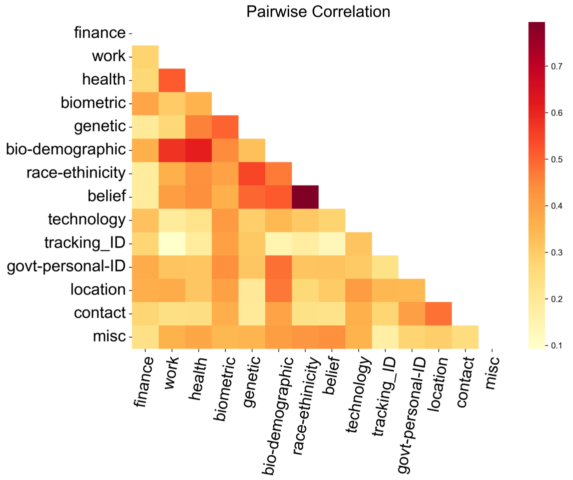

Figure 5: Pearson correlation between the occurrence of the PII types in the corpus.

We further compute the pair-wise correlations between the occurrence of PII types in each document to investigate the linear relationship between PII types. We show the results in **Figure 5**. We note that the correlations between all the PII types are always positive but differ for all the PII pairs. Although the correlations are positive, they are below 0.6, showing that no one pairing is dominant. The highest correlation is between Race/Ethnicity and Beliefs. Privacy laws that fall under this scenario include documents from 95 distinct countries and cover 16.29% of documents in the corpus. We also observed a high correlation between Genetic and Biometric PII types, which share a biological focus.

In **Figure 6** we show that all the PII types exhibit increase (i.e., the second derivative is positive) in frequency over time, but the rate of increase varies across the categories. For instance, we observe that for the "Biographical/Demographic", "Contact" and "health" increase has been rapid; however, for "Tracking IDs" the rate of increase is more sedate.

5.2. Topic Modeling

To explore the range of the topics covered in the privacy laws, we turn to algorithmic methods.

We leverage Latent Dirichlet Allocation (LDA), a probabilistic model, to extract latent semantic topics in the privacy laws (Blei et al., 2003). LDA model assumes that each document consists of several topics and that each topic is a distribution of words. Although every document in the Privacy Law Corpus concerns privacy, there are several dimensions to this topic. Therefore, we partition privacy laws into paragraphs to explore at a finer-grained level themes they contain.

Our privacy laws are stored in text files, but there are no discernible patterns to extract the paragraph structure precisely for a text document. Therefore, we use two newline characters (`\n\n`) as a proxy indicator of a new paragraph unit to extract the paragraphs from a document. It results in paragraphs with a vast range (1- 27,890 words) of length. To balance this range, we take a step further to divide the larger paragraphs into smaller paragraph units. We take a threshold of ten sentences and divide all the paragraphs we extracted in the previous step into the chunks of at most ten sentences. To filter out extremely small paragraphs, we remove all the paragraphs with less than nine words. With this technique, we are able to reduce the range to 9-1,133 words per paragraph. Each of these chunks forms a single input document unit for the LDA model.

We apply the following steps to preprocess the input text segments:

1. We tokenize all the segments into uni-grams,
2. We curate a custom list of stopwords (i.e. words carrying very little information). For our context, words like "article", "chapter", "number" provide insufficient information and, we include typical stop words such as "the", "is"" from gensim (gen, 2022). We then remove all the stopwords from the text segments,
3. We lemmatize all the tokens using WordNetLemmatizer (nlt, 2022),
4. We remove all the tokens with less than three characters, and
5. We filter out the tokens that occur less than 15 times and the ones present in more than 50% of the documents.

We generate a dictionary with the remaining tokens. Then the vector representation of each token is computed using TF-IDF (Sammut and I. Webb, 2011) and given to the LDA model. One hyperparameter of the LDA is the number of topics (k) to be considered. We experiment with six values for k (5, 6, 10, 12, 15, and 20) and by manual analysis, we find that the cohesiveness of the resulting clusters decreases with an increase in the k. We also experiment with a combination of uni-gram and bi-gram inputs and find that uni-gram results in a higher coherence score.

We manually interpret each output topic cluster by inspecting each topic's top ten relevant terms and the relevant documents. We get the best results for k equals six and show our results in **Table 5**. Out of these six clusters, four clusters show notable strong connections to the significant privacy concerns. These clusters cover Electronic Communication Service, Privacy and Data Protection, Government Regulations and Legislation, and Offenses/Penalty terms which are intuitively common industries for privacy concern. We also observe subtle similarities between the Offenses/Penalty terms and the Government Regulations and Legislation terms as the first describes the various aspects of penalty and prosecution, and the latter talks about the legal terms related to court and ministry. It is worth noting that these two topics suggest criminal laws. Given the broad inclusion criteria, we also include the criminal laws published by various jurisdictions that have sections devoted to privacy and data protection concerns.

For the remaining two clusters, the top relevant terms point to a combination of topics instead of a single topic. The topics are Legal Agreements and Labor and Employment Regulation. There is again some similarity between these two topics, where the first one is more in the context of contract and institution, while the other is more related to employment and work.

6. Conclusion

We introduce the Privacy Law Corpus, a collection of 1,040 official privacy laws from 183 jurisdictions around the world. We present our inclusion criteria for jurisdictions and privacy laws. To the best of our knowledge, this is the first of its kind of study. We contribute text in both English and the original version of the documents. By leveraging text analysis tools, we present a large-scale

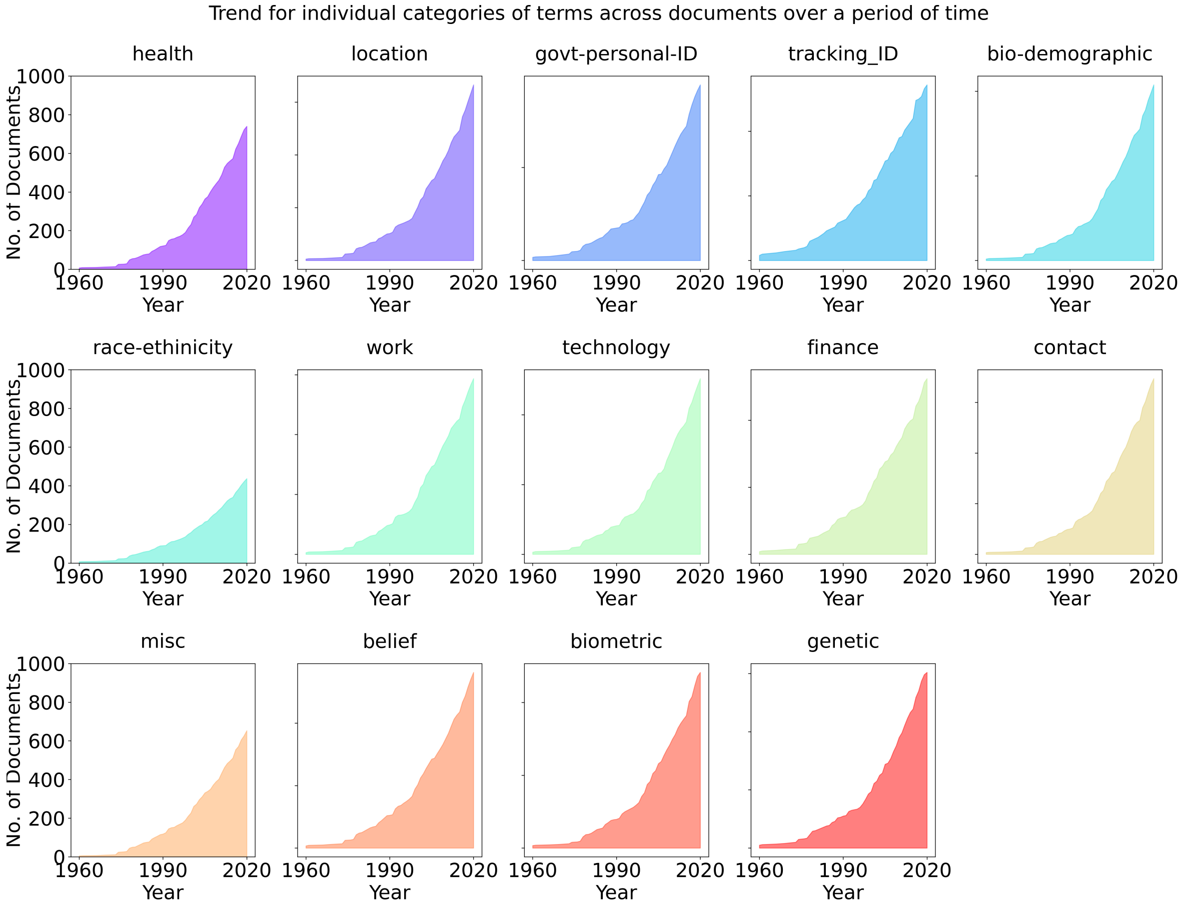

Figure 6: Trends for individual categories of terms across documents over a period of time

Legal Agreements	contract, right, company, institution, bank, person, payment, business, agreement, property
Electronic Communication Service	service, information, electronic, communication, system, document, public, network, national, security
Privacy and Data Protection	data, personal, information, protection, processing, person, subject, purpose, right, controller
Offenses and Penalty	year, person, fine, offense, imprisonment, penalty, amount, damage, case, criminal
Labor and Employment Regulation	work, employee, labor, employer, health, wage, employment, worker, medical, education
Government Regulations and Legislation	authority, code, state, provision, decision, regulation, member, minister, application, decree

Table 5: Terms in different categories.

empirical examination of the privacy laws and regulations published by governments to direct companies and organizations to pay attention to various aspects of consumers' privacy. We show how that attention has increased dramatically over time for some categories of Personally Identifiable Information (PII) more than others. We also see the signs of how this attention is distributed, resulting in only a few comprehensive privacy laws. In addition, we observe that certain PII types appear together more often than others. Some correlations are intuitive (e.g., Biometric & Genetic), while others (e.g., Race/Ethnicity & Beliefs) are relatively unexpected. Overall, the results provide previously absent nuances for claims that privacy is receiving increased attention and regulation. Finally, by releasing the corpus, we provide a basis for further work to examine privacy regulation on a global scale.

7. Future Directions[4]

There are several opportunities for future work in this space. Information retrieval techniques can be employed to develop a custom search engine and index all the privacy laws from the corpus. Such a service could enable easy access to the relevant documents and leverage the metadata to create filters for jurisdiction type, year of enactment, and more. With advancements in deep learning, conversational agent can be developed to answer the privacy law-related questions when requested.-Additionally, there exists opportunity to enhance our corpus by combining with data from various other sources such as news, politics and social media. For example, ParlaMint (CLARIN) is an EU-based corpora consisting of over 24 thousand parliamentary debate scripts and metadata. Combining such datasets with the Privacy Law Corpus can provide further insights into the evolution of privacy law.

8. Limitations[5]

We acknowledge limitations of this work. First, this work takes a perspective that is centered on the English language, to match the expertise of the authors. This perspective, along with the sheer volume of text, required the use of machine translation for some of the document collection process and for the text analysis. Second, the lack of international standards for what constitutes a “law”, “regulation”, “directive” or other government document means that creating a truly exhaustive collection of privacy laws is impractical. We mitigate that limitation through detailed collection rules and the use of online legal information resources, described earlier in the paper. Regardless of these limitations, we provide first-of-their-kind observations and a corpus that others can build upon to study the international privacy landscape.

9. Code and Corpus[6]

All text analysis code and the corpus is available at https://anonymous.4open.science/r/Privacy-Law-Corpus-CDFD/

10. Acknowledgements[7]

We acknowledge Prof. Graham Greenleaf’s itemization of privacy laws as an inspiration for our work: Global Tables of Data Privacy Laws and Bills (7th Ed, January 2021) Greenleaf (Greenleaf, 2021); Global Tables of Data Privacy Laws and Bills (6th Ed, January 2019) Greenleaf (Greenleaf, 2019). The present project began with collecting the set of laws that his tables refer to, continuing outward from that set. We are also grateful to Prof. Greenleaf for his correspondence while building this corpus. Additionally, the work reported herein was supported in part by the National Science Foundation under Grant #CNS-1914444, #CNS-1914486, and #CNS-1914446. .

11. Bibliographical References

2022. Data guidance. https://www.dataguidance.com. Accessed: May 10, 2023.

2022. Data privacy act (2012), philippines. https://www.privacy.gov.ph/dataprivacy-act/.

2022. DLA Piper. https://www.dlapiper.com/en-us/. Accessed: May 10, 2023.

2022. European commission. https://commission.europa.eu/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_en. Accessed on September 23, 2023.

2022. Source code for gensim.parsing.preprocessing. https://tedboy.github.io/nlps/_modules/gensim/parsing/preprocessing.html. Accessed on September 24, 2023.

2022. Source code for nltk.stem.wordnet. https://www.nltk.org/_modules/nltk/stem/wordnet.html. Accessed on September 24, 2023.

2022. United Nations. https://www.un.org/en/about-us/member-states. Accessed: May 10, 2023.

2022. Yandex. https://yandex.com/company/. Accessed on May 30, 2023.

2023. 2016 turkish coup d’état attempt. Page Version ID: 1154360104.

U.S. National Archives and Records Administration. 2022. Controlled unclassified information (cui).

David M Blei, Andrew Y NG, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

[4] Per the LREC template, this section is exempt from the page limit.

[5] Per the LREC template, this section is exempt from the page limit.

[6] Per the LREC template, this section is exempt from the page limit.

[7] Per the LREC template, this section is exempt from page limits.

CLARIN. ParlaMint. Accessed on March 17th, 2024.

European Commission. 2021. What's personal data? https://commission.europa.eu/law/law-topic/data-protection/reform/what-personal-data_en. Accessed on September 24, 2023.

Michael A. Elliott. 2011. The institutional expansion of human rights, 1863–2003: A comprehensive dataset of international instruments. *Journal of Peace Research*, 48(4):537–546.

World Privacy Forum. 2021. National IDs of various countries. https://www.worldprivacyforum.org/2021/10/national-ids-and-biometrics/. Accessed on September 24, 2023.

FTC. Federal trade commission.

Graham Greenleaf. 2014. *Asian Data Privacy Laws: Trade Human Rights Perspectives*. OUP Oxford.

Graham Greenleaf. 2019. *Global Data Privacy Laws 2019: 132 National Laws Many Bills*. The Privacy Laws Business International Report.

Graham Greenleaf. 2021. Global tables of data privacy laws and bills (january 2021).

Mitra Bokaei Hosseini, Travis D. Breaux, Rocky Slavin, Jianwei Niu, and Xiaoyin Wang. 2021. Analyzing privacy policies through syntax-driven semantic analysis of information types. *Information and Software Technology*, 138:106608.

Guiraude Lame. 2005. Using nlp techniques to identify legal ontology components: concepts and relations. *Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications*, pages 169–184.

Claire Bazy Malaurie, Sarah Cleveland, Regina Kiener, Hanna Suchocka, Kaarlo Tuori, and Jan Velaers. 2016. Opinion on emergency decree laws nos 667-676 adopted following the failed coup of 15 july 2016 in turkey. Adopted by the Venice Commission at its 109th Plenary Session (9-10 December 2016).

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripa Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation. *arXiv preprint arXiv:2105.13562*.

Julián Moreno-Schneider, Georg Rehm, Elena Montiel-Ponsoda, Víctor Rodriguez-Doncel, Artem Revenko, Sotirios Karampatakis, Maria Khvalchik, Christian Sageder, Jorge Gracia, and Filippo Maganza. 2020. Orchestrating nlp services for the legal domain. *arXiv preprint arXiv:2003.12900*.

Shunji Mori, Hirobumi Nishida, and Hiromitsu Yamada. 1999. *Optical Character Recognition*, 1st edition. John Wiley & Sons, Inc., Indianapolis, Indiana.

NIST. 2022. National institute of standards and technology. https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-122.pdf.

Pratum. 2020. The national impact of ccpa. https://pratum.com/blog/425-the-national-impact-of-ccpa. Accessed: May 10, 2023.

Abhilasha Ravichander, Alan W. Black, Thomas Norton, Shomir Wilson, and Norman Sadeh. 2021. Breaking down walls of text: How can nlp benefit consumer privacy? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, volume 1.

Livio Robaldo, Serena Villata, Adam Wyner, and Matthias Grabmair. 2019. Introduction for artificial intelligence and law: special issue "natural language processing for legal texts".

Claude Sammut and Geoffrey I. Webb. 2011. TF–IDF. *Encyclopedia of Machine Learning. Springer, Boston, MA*, pages 986–987.

Yan Shvartzshanider, Ananth Balashankar, Thomas Wies, and Lakshminarayanan Subramanian. 2018. Recipe: Applying open domain question answering to privacy policies. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 71–77.

The Apache Tika Development Team. 2021. Apache tika. https://tika.apache.org. Accessed: May 11, 2023.

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, et al. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340.

Shomir Wilson, Florian Schaub, Frederick Liu, Kanthashree Mysore Sathyendra, Daniel Smullen, Sebastian Zimmeck, Rohan Ramanath, Peter Story, Fei Liu, Norman Sadeh, et al. 2018. Analyzing privacy policies at scale: From crowdsourcing to automated annotations. *ACM Transactions on the Web (TWEB)*, 13(1):1–29.

Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R. Reidenberg, N. Cameron Russell, and Norman Sadeh. 2019. Maps: Scaling privacy compliance analysis to a million apps. *Proc. Priv. Enhancing Tech.*, 2019:66.

12. Appendices

Appendix A. Jurisdiction Inclusion Criteria

We include a jurisdiction if it satisfies one of the two requirements: (i) it is a country recognized as either a member or observer state of the United Nations by at least one other member state as of 2020, and (ii) a jurisdiction falls into the following special categories: (a) self-governing British Overseas Territories (Bermuda, Gibraltar, Cayman Islands), (b) crown dependencies (Guernsey, Jersy, Isle of Man), (c) Chinese Special Economic Regions (Macau and Hong Kong), (d) Qatar economic free zones (Qatar Financial Centre), (e) United Arab Emirates economic free zones (Abu Dhabi Global Market, Dubai International Financial Centre, Dubai Healthcare City), and (f) the states which are not recognized as UN members or observers (The Republic of China (Taiwan) and Kosovo). We also include one US state(California), due to its significance and weight in defining privacy legislation that impacts the entire US economic system (Pratum, 2020).

Appendix B. Document inclusion criteria - Based on document type

Each document must meet at least one of the following inclusion criteria:

- The document is legally enforceable (or once-enforceable and now defunct, or assumed to be enforceable upon some future date of effect), which is promulgated in a complete state to the general public for the purposes of awareness of the law and enforcement if it is in force, which may include laws and regulations

- The document contains rules, clarification, or similar resources directed towards lawmakers or law enforcement for the purposes of enforcing the aforementioned document.

- The document contains a non-enforceable list of guidelines, which serve as official guidance directed towards the general public, or specific sectors of the public, for the purposes of advising them on how to comply with a document of another type.

Appendix C. Document inclusion criteria - Based on source type

The documents must meet any one of the following inclusion criteria for document source:

- The document contains more content than a notification containing some update regarding the legal status of another document. A decree that says only that a different law is now in effect, providing no further guidance or substance, is excluded.

- The document is released by a government entity, such as (but not limited to) an executive order released by a president, a law passed by a congress or parliament, or a set of rules released by a government agency. Documents released by non-government entities, such as rules released by corporations and non-profit organizations for the internal governance of data privacy, guidelines released for the general public, and others, are excluded.

- The document is released to the general public with the intent of circulating the document in its current, complete state for the purposes of understanding or enforcement of the document. Such circulation resources may include government websites and legal journals. This implies that the following types of documents are excluded.
  - Private documents are not meant for such release to the public.
  - Rules that describe internal procedures not directly relevant to the privacy laws and concepts in question, such as documents that merely describe which agencies or positions are charged with particular enforcement duties, are not included in this corpus. This is because these documents do not provide meaningful context into how the meaning of law itself is interpreted and enforced.
  - Activity reports of government agencies, meant primarily for internal review and as a resource regarding the state of enforcement. Because of their conceptual removal from the types of documents of interest to the researchers.
  - Strategy and action plans designed for internal use by enforcement agencies.

  - Enforcement decisions and records of fines. This is because they are notices aimed towards the specific audience of a given punished entity, without a desired audience of the general civic public.

- While future versions of the document may be released with changes, the document is released within its given form with the understanding that this form is immutable and is to be understood as-is until further documents are released to update it. This implies that the following types of documents are excluded

  - Bills and similarly unfinished documents released in various drafts for the purposes of transient public forum discussion.
  - Forms, software tools, and other tools that require active constituent participation for effective use. As the form of these artifacts extends beyond the static, immutable document states that we wish to analyze here.

- The documents are promulgated in their included region by or before December 31, 2020. We set that date significantly in the past to promote higher recall in the final years of the corpus, recognizing that documents from some jurisdictions are not immediately available online.

Appendix D. Excluded document types

These document types exclude case law, which establishes legal precedents through individual court decisions. Although such cases are valuable pieces of information and form precedents for decisions regarding compliance with laws related to privacy, they neither form an explicit legal directive or instruction nor a document explicitly instructing the reader about how to enforce or comply with such instructions. Additionally, due to the overwhelming scope and limited resources for acquiring case law notices or summaries globally, case law is categorically excluded from this work.

This corpus also excludes discussions of legal rationale unaccompanied by content that matches the aforementioned document types. Much like case law, discussions and arguments explaining the rationale behind a legal directive are a malleable resource that can be used to understand the application of the law. However, we exclude them because such documents also do not provide any direct instruction or guidance to the reader and, instead summarize lawmakers' theoretical decisions.

The final notable type of document excluded from this work is national constitutions, which provide established principles with significance both in their own right as legal documents and as a potent precedent for other laws developed in the country. We categorically exclude national constitutions because, in the overwhelming majority of cases, allusions to a right to privacy in a constitutional document were found to lack actionable details regarding expectations, instructions, or enforcement. Thus, although such mentions within national constitutions may act as a guiding principle in the development of subsequent legal documents regarding privacy, we find that these constitutional documents do not provide enough instruction to lawmakers, enforcers, or citizens regarding privacy to be a meaningful and effective part for our corpus.

Appendix E. Legal Enforceability of Corpus Documents

In the corpus, we mark each English translated document with whether the translation is completed by a human government translator, a 3rd-party government translator or by our own machine translation. However, levels of legal enforceability of each type of document vary widely among countries and individual instances. For example, there are English translations of laws released by both third-party groups and government resources that proclaim that they are for informational use only and that the law is only legally enforceable in the original non-English language. In contrast, some jurisdictions appear to provide their laws in multiple languages but fail to specify which version of the document is legally enforceable. The translations generated by the 3rd-party groups (e.g., private law firms) are less likely to be strictly legally enforceable than documents sourced directly from government websites. As the clarifications of legal validity or non-validity (or the absence thereof) vary wildly among the documents from all the different categories of sources, we are not able to definitively mark each document as technically legally enforceable in its current state. Thus, when we mark a particular document as "in effect", this is to say that the original law is "in effect", but this does not guarantee the legal accuracy and permissibility of all of the permutations of each document we provide within the corpus. These are only for informational and research purposes and cannot be assumed to be legally viable forms of these laws, regulations, and recommendations. This limits the applicability of this corpus as an up-to-date legal tool for use by legal counsel or by end-users seeking to guarantee their compliance with privacy laws, regulations, and recommendations.

Appendix F: PII Types and List of Keywords

There are 14 different categories of Personally

Identifiable Information (PII) that are discussed in this paper. A complete list of all keywords that were used is included in the following table

PII Types	Example Keywords
Finance	Bank Account Number, Credit Card, Credit Card Number, Debit Card, Debit Card Number, Economic Situation, Financial Account, Financial Information
Work	EIN, Employee Information, Employee Emergency Contact, Employment, Employment History, Professional, Salary
Health	AIDS, Biological, Breastfeeding, Cancer, Childbirth, Chiropractic, Dentistry, Diagnosis, Disability, Disease Prevention, Fitness, Health, Health Condition, Health Insurance Information, Health Promotion, Health Status, Healthcare, Healthcare Information, Healthcare Provider, HIV, Medical Condition, Medical Diagnosis, Medical History, Medical Information, Medical Record, Medical Treatment, Medicine, Mental, Mental Condition, Mental Health, Nursing, Occupational Therapy, Olfactory, Optometry, Patient, Pharmacy, Physical Health, Pregnancy, Protected Health Information, Sleep
Biometric	Biometric, Biometric Data, Biometric Information, Face, Faceprint, Facial Recognition, Fingerprint, Gait, Iris, Retina, Thermal, Unique Biometric, Voice, Voiceprint
Genetic	Biological Characteristic, Deoxyribonucleic Acid, DNA, Eye Color, Genetic, Genetic Data, Genetic Information
Bio./Demographic	Age, Birth, Cell Phone Number, Date of Birth, Death, Divorce, Education, Family, First Name, Full Name, Gender, Gender Expression, Gender Identity, Height, Language, Last Name, Marital Status, Marriage, Middle Name, Name, National Origin, Personal Data, Place of Birth, Real Name, Sex, Sexual Orientation, Surname
Race/Ethnicity	Ancestry, Color, Ethnic, Ethnic Origin, Ethnicity, Race, Racial, Racial Origin
Beliefs	Religion, Religious, Philosophical, Political, Political Belief
Technology	Access Code, Audio, Automated Processing, Browsing History, Dark Pattern, Network Activity, Password, User ID, Video
Tracking	Beacons, Device Identifier, Internet Activity, Internet Protocol Address, Internet Protocol IP Address, IP Address, MAC Address, Online Identifier, Persistent Identifier
Govt./Personal IDs	Aadhar, Citizenship, Citizenship Status, Customer Number, Documento Nacional de Identidad, Driver License, Identification Card Number, Identification Number, Insurance Policy Number, Military Status, National ID, Passport Number, Personal Identification Number, SSN, Social Security Number, State Identification Card, Taxpayer ID
Location	Address, Coordinates, Geographic Area, Geolocation, Geolocation Data, Geolocation Information, Home Address, Latitude, Locate, Location, Longitude, Postal Address, Precise Geolocation, Radius, Specific Location, Work Address
Contact	Contact, Electronic Mail Address, Email Address, Fax Number, Mobile Number, Mobile Phone Number, Phone Number, Phone No., Telephone Number
Misc.	Alias, Decree, Email Content, Image, Investigation Report, Notes, Union Membership

Table 6: PII types along with its sample keywords.