

Regulated Personal Data Type Recognition Task Definition

Version 0.4
2020-03-21

This document was used to create the dataset in the following paper.

Duc Bui, Kang G. Shin, Jong-Min Choi, and Junbum Shin, "Automated Extraction and Presentation of Data Practices in Privacy Policies," Proceedings on Privacy Enhancing Technologies, 2021.

Goal	1
High-level Principles	1
Annotation Guidelines	2
How to Revise Existing OPP-115 Annotations	2
How to Determine Text Spans for Personal Data Types	3
Prefixes	3
Suffixes, wh- and if-clauses and relative clauses	5
Suffixes - prepositional phrases	6
List item separation	6
Other cases	9





Goal

We aim to extract a list of personal data types that are collected/used, or shared in online privacy policy documents. This guideline document helps the extracted noun phrases consistently among different annotators.

High-level Principles

1. Extract a list of *noun phrases* which are the regulated personal data types.
2. Mark the words so that **when the users read only the extracted words, they still can make sense of them**. That means the user should be able to understand the extracted data types in the list without reading their context in the document.
3. Prefer more specific data types than general data types.
4. Include words into the data types only when it makes the data type more specific.
5. Prefer shorter than longer text spans to reduce the time needed to read.
6. Automated correction and automated pre-annotation have limitations due to the imperfection of the algorithms. Therefore, the annotators should pay attention to avoid the biases introduced by the automated methods.

In other words, mark the words that you expect to be highlighted when asking the question: "What information is collected/used/shared by the service?"

Highlight color	Label
	COLLECT
	NOT_COLLECT
	SHARE
	NOT_SHARE

Annotation Guidelines

How to Revise Existing OPP-115 Annotations

Principles:

- Extract noun phrases from the existing annotations.
- Rely on the existing annotations because they were created by experts.
- If you may add/delete the existing annotations, if the meaning is clear from the context. If you have any doubt, especially in hard cases, just keep the existing annotations.

Examples of annotation refinement:

1. Add a prefix which provides more details (rule 2).

we share non-personally identifiable information with third parties → we share non-personally identifiable information with third parties
[add "non-personally identifiable" adjective]

2. Reduce prepositional prefixes (rule 2).

advertisers may place cookies in your computer in order to collect certain information about your use of the service → advertisers may place cookies in your computer in order to collect certain information about your use of the service
[remove "in order to"]

3. Decompose long annotations into smaller ones (rule 7).

We collect **your information such as your location and demographics** to provide you with personalized services. → We collect **your information** such as **your location** and **demographics** to provide you with personalized services

How to Determine Text Spans for Personal Data Types

This section describes rules to annotate the documents to produce consistent text spans.

List item separation

1. Inseparable lists: mark all the list if it indicates different data types, to make a complete phrase for personal data types (although it may contain more than 1 personal data type). For example, "adj1 and adj2 information" pattern.

first and last name

Time and date of calls

we measure **traffic and usage trends**

we collect **OS type and version**.

We collect **content of your reviews and e-mails you send to us**.
[This list is not separable because of the phrase "you send to us"]

your general location (**your country** or **city area**)

the **address of the external or internal page that referred you**
["external or internal" both are related to to the "page"]

We place a simple cookie to remember **your playback preferences** including **shuffle** and **NSFW/SFW**.
["NSFW/SFW" is not separated because they are used as if they are a personal data type]

We collect **information about the service that you use and how you use them, like when you watch a video on YouTube, visit a website that users our advertising services, or view and interact with our ads and content**.

we use technologies to collect your **information on nearby devices, WiFi access points and cell towers**.

The Google Analytics products helps businesses and site owners analyze the **traffics to their websites and apps**.
["websites and apps" is considered as one]

referring/exit pages and URLs

demographic and marketing analyses of users of the service, and their subscribing and purchasing patterns.

people (with addresses and phone numbers) listed in 1-Click settings

e - mail addresses of your friends and other people

Paypal may send us a receipt with your name and email address

software and hardware attributes

[it may be hard for user to understand extracted "software" and "hardware attributes" separately]

OS type and version

shipping or billing address

["shipping" and "billing" here act like adjectives]

home and email address

your student or faculty identification number

[the "identification number" is not general]

Data logged via PS3 dynamic in - game advertising or some portion of it

[because of "it" refers to "Data logged..."]

interest in , and collect aggregate information on , our websites

information that you provide us , or that we collect about you

information you enter on our Websites or that you give us in any other way

login and password information

2. Separable lists: mark individual items. For example, "noun1 and noun2 information" or "information of noun1 and noun2" patterns.

We collect information about your device, including IP address, web browser name and OS version.

We do not give that business your name and address. ["address" alone can be understood as the user's address and still makes sense]

You need to create a profile which may include your name and photo.

SMS routing information and types of calls.

Members will participate in surveys, polls or discussions about their readership of The New York Times, their household / personal characteristics and their purchase behavior.

[decompose the long list "surveys, polls, ... purchase behavior" into smaller data types.

We collect personal information about the computer, the mobile device or other device you use to access the service.

information, reports and analysis about the usage, browsing patterns of our users.

Web beacons allow us to know if a certain page was visited, an email was opened, or if ad banners on your websites are effective.

search term and search result information from some searches conducted through the Web search features

your credit card number and zip code when you buy a ticket

[select 2 text spans separated by "and" because the users expect 2 pieces of personal information]

personal description and photograph in Your Profile

We receive updated delivery and address information.

[Mark the "and" because this sentence should be "updated delivery information" and "updated address information"]

We collect your computer and connection information.

[Mark the "and" because of the word "information", it does not make sense when extracting only "we collect your computer"]

We release account and other personal information.

[This list can be written in full "account information" and "other personal information"]

Google Analytics tells us aggregate usage and traffic information.

[mark all due to "information" word]

computer and connection information

demographic and preference information

["demographic" is a noun here, so can understandable without context]

Other cases

3. Do not annotate tracking technologies which are not the collected/used/shared data types, such as **cookies, pixel tags, pixel trackers, local storage, browser web storage, application data caches, databases, server logs, and Google Analytics**. We should choose whole phrases

which include the collection technology if it specifies the information we need to take, e.g., the cookies are the data that the service collects.

We collect some information. This includes cookies that my uniquely identify your browser or your Google Account.

We use various technologies to determine location including IP address, GPS, and other sensors that may, for example, provide use with information on nearby devices, WiFi access points and cell towers.

4. Mark all words which are tokenized with a dash in the middle such as:

e - mail

plug - in

device - specific information

Wi - Fi

add - ons

opt - out

opt - in

5. Mark the examples/objects that the users use. We consider the service will collect information about the objects.

information regarding your computer or other device used to access our service (such as gaming systems, smart TVs, mobile devices, and set top boxes).

[imply the service will collect information about gaming systems, smart TVs, mobile devices, set top boxes]

6. Mark the abbreviations in parentheses so that the extracted words are meaningful (otherwise, they are fragments). This is because we do not consider non-contiguous annotations.

We collect the internet protocol (IP) address used to connect your computer to the Internet.

7. Do not select pronouns like "it" because it is hard to understand without context.

They have access to personal information needed to perform their functions, but may not use it for other purposes.

[Do not select "it" on the second clause]

Note on distinguishing collection/use vs. sharing

Sharing does not always imply Collection/Use since the service can hire a third party to analyze the data of the user (like using an embedded Google AdSense script), while the service itself does not collect the data.

[end of guidelines]