

## Data enumeration process for Table 3 (APPG Extent of Data Recognised)

In this document, we delve deeper into the analytical process of identifying and categorizing various data elements and practices recognized by Automated Privacy Policy Generators (APPGs). The objective is to discern the scope and depth of data utilization, and the practices APPGs identify to support the formation of privacy policies.

Data Enumeration Process:

### 1. Identification of Major Aspects

Initially, the investigation categorizes the data recognition capabilities of APPGs into four broad categories:

**App's Basic Information:** This includes the foundational details about the app and its developer.

**Users' Personal Information:** This encompasses both general and sensitive personal data collected from the end-users.

**Device Permissions:** This section details the various permissions an app may request to access different functionalities on an end-user's device.

**Third-party Services:** This includes the recognition of data use and sharing with selected third-party services.

### 2. Traversal and Count for Question and Option

To evaluate the recognition capabilities of APPGs, we intend to collect all potential questions and options that users might face while using. As some questions will only be popped up after ticking some prerequisites questions, to guarantee the completeness, we follow the "Depth-First Searching (DFS)" strategy to conduct the exploration. For example, the Question 2 will only appear in the next page if we select "Yes" for Question 1 in the red bounding box.

**User Location**

Do you have users in the United States?  
Select "Yes" if you want to be CCPA compliant.

☒ Yes  
☐ No

Do you have users in the EU, UK, Switzerland, Iceland, Liechtenstein, or Norway? ?

☒ Yes  
☐ No

Do you have users in Canada?

☒ Yes  
☐ No

Figure 1 Question 1 in red bounding box

**California (USA) Users**

Do you want your privacy policy to be California Consumer Privacy Act (CCPA) compliant?  
If your for-profit business collects the personal information of users located in California (or may do so in the future), you may be required to comply with the CCPA. Check the FAQ on the right to see if you meet the requirements.

☐ Yes, I wish to be CCPA compliant.

Figure 2 Question 2

Here is another example, the question about personal information selection will appear only “Yes” is selected for the first question.

The image displays two screenshots of a web form titled "Privacy Policy".

**Left Screenshot:** The form is titled "Privacy Policy" and "Sensitive Personal Information Collected". It asks "Do you collect sensitive information?" with a subtext "If you are not sure, select 'Yes' to see examples on sensitive information." The "Yes" radio button is selected and highlighted with a red box. Below the question, there is a "BACK" button and a "SAVE & NEXT" button.

**Right Screenshot:** This screenshot shows the same form but at a different step. It asks "Please select the sensitive personal information you collect:" with a subtext "Generally, personal information categorized as sensitive must be treated with more care and caution." Below this, there is a list of checkboxes for various types of sensitive information: Health data, Financial data, Genetic data, Biometric data, Data about a person's sex life or sexual orientation, Information revealing race or ethnic origin, Information revealing political opinions, Information revealing religious or philosophical beliefs, Information revealing trade union membership, Credit worthiness data, Student data, and Social security numbers or other government identifiers. There is also a text input field for "Add your own" with a "+ ADD" button. At the bottom, there is a note: "Sensitive categories of personal information must be treated with additional care because of the risk imposed on the data subject." Below this, there is a "BACK" button and a "SAVE & NEXT" button.

During this process, we also counted the number of questions and recorded their types, which comes to the **Statistic summary** section in the Table 3. The “Minimum questions” denote the shallowest steps of our DFS strategy, and the “Maximum questions” denote the total unique nodes (questions) during our DFS traversal.

### 3. Data Classification

For each question and option mentioned in the previous step, two annotators independently judged whether it explicitly pertains to a specific data type, permission, or third-party service, then classified as “recognised” (\CIRCLE), otherwise “absent” (\Circle).

As it involves intensive manual work, to avoid the effect caused by potential human error, we employ the same strategy as introduced in the previous section.

Both annotators labelled “recognised” for 148 items and “absent” for 149 items. 6 items are labelled as “recognised” only by annotator A, and other 7 items are labelled as “recognised” only by annotator B. Thus, the Cohen’s Kappa  $\kappa = 0.92$  for the initial manual labelling, which is an almost perfect level of agreement. For those disagreements, they discussed and agreed on the same answer.

