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.

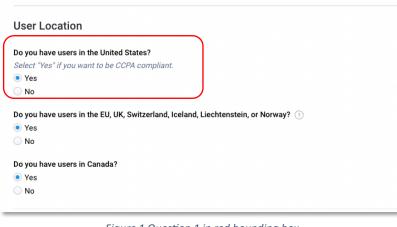


Figure 1 Question 1 in red bounding box



Figure 2 Question 2

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

	Privacy Policy			
	Sensitive Personal Information Collected			
	Do you collect sensitive information?			
	If you are not sure, select "Yes" to see examples on sensitive information. •) Yes No			
	Please select the sensitive personal information you collect:			
	Generally, personal information categorized as sensitive must be treated with more care and caution. 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			
Dalaran Dallara	Social security numbers or other government identifiers			
Privacy Policy	Add your own			
Sensitive Personal Information Collected	+ ADD			
Do you collect sensitive information? If you are not sere, select "feet" to see examples on sensitive information. Yes NO	Sensitive categories of personal information must be treated with additional care because of the risk imposed on the data subject.			
BACK SAVE 6 NEXT	BACK SAVE & NEXT			

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 κ = 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.