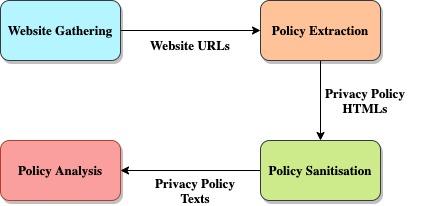
# 3. Design

## 3.1. Design overview

The architecture of the project has four main components, which operate independently from one another. The aim was that each component could be used separately in future works.



**Figure 1. Architecture Design Diagram**

## 3.2. Website Gathering

The website gathering component of the programme aims at generating a list of website urls that can be easily accessed and used for future analysis.

In order to achieve that goal, one option was to create a scraper that would harvest URLs from an existing list available online. Another option was to retrieve a list of websites from an API meaning that the list is refreshed whenever a new API request is made. The later option was adopted, because it helps ensure that the data collected can be more easily updated, thus making the programme more robust.

All the URLs collected are then formatted correctly so that each website on the list can be accessed directly, either manually or with an automated tool.

## 3.3. Policy Extraction

The policy extraction component is designed to extract a privacy policy page from a specific website, and return the content of the page. Furthermore, the programme aims at mimicking how humans would behave if they were looking for a website’s privacy page, in order to identify potential strengths or weaknesses in the way organisations deal with privacy online. This will be further discussed in the Evaluation chapter of this report.

A web scraper was adopted as the solution to access the websites, find their privacy policy pages and extract the content of the page. A web scraper is an automated software application that is used to extract data from targeted web sources [40]. This tool can be viewed as a robot mimicking the functions of a human by interacting with websites and extracting data stored in it [41][42], which is precisely what the policy extraction component of this project is attempting.

A feature provided by a scraper when extracting data from web sites is that the data can be easily kept up-to-date. This is a key advantage when dealing with privacy policies, because these documents can be modified regularly. To make the policy extraction component reusable over time, the tool should be reusable to get updated data for new analysis.

Choosing to design the policy extractor as a web scraper also makes the tool scalable, because it can theoretically be used on any given website, assuming that the site’s URL is valid and that access to the site is not blocked by any security protocol.

The web scraper is designed to find specific words or links in a page that contain the word “privacy” similarly to what a human do if manually searching for the link to a website’s privacy policy, or using the “ctrl + f” browser shortcut. The scraper can also be customised to look for keywords in other languages.

The policy extractor is designed to store the content of each privacy policy in a file on the local machine, in order to be used later.

## 3.4. Policy Sanitisation

Expectedly, data collected from a web page will need sanitisation before being used for any sort of analysis. This component of the programme is designed to harmonise the data, ensuring that all the policies are in the same format and removing certain elements that might cause errors when analysing the data are removed. The aim is to obtain privacy policies in a format that is both computer and human-readable, to be adaptable to different types of policy analysis. The policy sanitisation is treated as an individual component, but it could also be contemplated as the final part of the scraper, cleaning the data before storing it in a file, or as the pre-processing step in the data analysis.

## 3.5. Policy Analysis

The last component of this project’s architecture is the analysis of privacy policies, which aims at finding interesting information relating the legal basis for data processing used by companies for compliance with Article 6 of the GDPR.

Different methods were explored to achieve this aim. The first approach involved manual review and a heuristic search of keywords connected to Article 7. Another method to find information in connection to a specific topic, in this case legal grounds for data processing, is to use topic modeling on the whole set of documents. Topic modeling is an unsupervised machine learning method which attempts to extract the most probable distribution of words into topics through an iterative generative process which terminates upon convergence [16]. This is particularly useful in large datasets where it is difficult to extract topics manually.

The design of this project’s policy analysis component is a topic model which can analyse the collection of privacy policies and discover abstract topics. The resulting topics are then inspected to find pre-selected heuristics associated with the vocabulary of the legal grounds for data processing.