Web scraping policy

Our policy for web scraping on food.gov

Background

Use of alternative data sources is vital for the FSA to achieve its objective of protecting public health from risks which may arise in connection with the consumption of food (including risks caused by the way in which it is produced or supplied) and otherwise protecting the interests of consumers in relation to food. Web Scraping is one of these alternative data sources often used to supplement existing data sources. Therefore, the purpose of this policy is to ensure Web Scraping activities are conducted transparently, ethically and with respect to all relevant legislation.

This policy sets out the practices and procedures that the Food Standards Agency will follow when carrying out web scraping. Web scraping is defined as the automatic collection of data retrieved from the internet, typically from HTML, using a piece of software or program.

Scope

This policy sets out the practices and procedures that Food Standards Agency employees, temporary and casual staff, and any contractor, researcher or academic partner carrying out work for the agency or acting on our behalf must follow when carrying out web scraping.

When assessing external organisations who may scrape on behalf of the FSA, either via business as usual activities or on an ad-hoc basis, this policy will be used to guide assurance activities. This is to ensure appropriate governance structures are in place when gathering and processing data gathered via scraping activities.

Note that this document does not cover the use of Application Programming Interfaces (API's) and applies only to the Food Standards Agency.

Next review

We aim to keep all guidance up to date and undertake regular reviews to ensure this guidance remains relevant. The next scheduled review date for this guidance is 15 April 2022.

Compliance

All FSA staff must comply with this policy in any project involving scraping data from the web. Failure to comply may result in disciplinary action in line with the organisation's discipline policy.

Policy statement

The FSA is carrying out web-scraping activity for the purposes of its functions set out in the Food Standards Act 1999.

The FSA's legal basis for such webscraping activity are the following:

Sections 8, 10, 11 and 21(1) of the Food Standard Act 1999 give the FSA the powers of obtaining and compling information, by carrying out observations, about matters relating to food safety or the interests of consumers in relation to food, including but not limited to food sources, food premises, food businesses or commercial operations being carried out with respect to food.

In carrying out the webscraping activity described herein the FSA is lawfully exercising its statutory powers and is acting within its statutory authority. Section 19 of the Food Standards Act 1999 also allows FSA to share information where to do so would be in the public interest.

To the extent that the information which is scraped identifies living individuals the scraping activity involves the processing of some personal data. The FSA will process such data in accordance with the provisions of the EU General Data Protection Regulation 2016/679 (GDPR) and the UK Data Protection Act 2018 (DPA)

The FSA believes that the processing is necessary within the public task function of the FSA (GDPR Article 6.1(e)) and, to the extent criminal data is relevant, necessary for the FSA as a competent authority for law enforcement purposes in the public interest and/or for legal claims (GDPR Article 9.2(f) and (g) and DPA Schedule 2 Part 1 and DPA Schedule 2 Part 1, section 5 and Schedule 7) or has likely been placed in the public domain by the individuals concerned (GDPR Article 9.2(e)).

The FSA will adopt the following principles to guide our web scraping activities:

- Seek to minimise the burden on scraped sites
- Abide by all applicable legislation and monitor the evolving legal and ethical situation
- Protect all personal data

Policy details

Seek to minimise the burden on scraped sites

Excessive use of web scraping can be burdensome, and overuse on a single webpage may have a detrimental impact to the operation of the website. To avoid such scenarios, we will follow these practices, where applicable, and always act within the limits of our statutory functions:

- Assess other data collection methods, such as API's, before considering the use of web scraping
- Respect scraping limitations enforced via documents like the Robots Exclusion Protocol (Robots.txt) and Site Terms and Conditions
- Delay accessing pages on the same domain
- Adding idle time between requests
- Limiting the depth of crawl within the given domain
- When scraping multiple domains, parallelising the crawling operation to minimise successive requests to the same domain
- Scraping at a time of day when the website is unlikely to be experiencing heavy traffic - for example, early in the morning or night
- Optimising the web scraping strategy to minimise volumes of requests to domains
- Only collect the parts of pages required for the initial purpose
- Include a User Agent String to differentiate FSA from other users

Please note that this list of best practices is non-exhaustive and the FSA will strive to maintain a high-standard of web-scraping by monitoring and implementing best practices.

Abide by applicable legislation and monitor the evolving legal and ethical situation

The FSA understands that the legal situation surrounding web scraping is complex and ever evolving; and there are relatively few relevant legal precedents for conducting such activities. The FSA will continue to monitor the legal situation as it evolves and amend our approach accordingly. Likewise, the FSA recognises the use of web scraping and the processing of data collected in this manner may give rise to a variety of ethical issues and challenges. Efforts are currently underway to develop a practical framework for considering and mitigating potential ethical risks both directly and indirectly attributed to web scraping.

Those who wish to undertake web scraping activities must document and justify the rationale, the legal and ethical reasoning and the benefits of conducting the web scraping. This rationale will be used as the basis for approval from the Information Governance Board and will be scrutinised during the assessment phase to ensure continued alignment with applicable legislation. See section 9 for more details on this.

Protect all personal data

The FSA is fully commited to compliance with the Data Protection Act 2018 and the General Data Protection Regulation and ensuring that personal data is used lawfully, fairly and transparently. Data protection is of great importance to the FSA, not only because it is critical to the work of our organisation, but also because it ensures we protect individual privacy and maintain public confidence in the agency.

Roles and responsibilities

Role	Responsibility
FSA Staff conducting web scraping activities	• Complying with the web scraping policy • Adhering to the agreed web scraping process established across the agency
Knowledge and Information Management and Security (KIMS)	• Disseminating the policy and making staff in all divisions aware of it as part of the overall FSA Information Governance Framework
Information Governance Board	• Receiving and approving proposed web scraping activities from the FSA • Providing a justification for disapproved web scraping activities
Legal services	• Provide advice on current legal issues and evolving landscape if required

Governance

Policy owner	Knowledge and Information Management and Security (KIMS)
Policy approval	Information Governance Board

Compliance monitoring	Knowledge and Information Management and Security and Head of Divisions
Review and amends	Data and Knowledge and Information Management and Security teams

Accompanying guidance

Web scraping steps

Web scraping request

- FSA business units wishing to scrape should contact the Heads of Data and KIMS as soon as possible after identifying the requirement
- A completed web scraping request document must also be attached to the correspondence for review
- Please note that the FSA may conduct a feasibility scrape on your website. No data will be stored by the FSA during this time. The FSA will ensure that this process remains in line with section 6 of this document, however will not be subject to evaluation by the Information Governance Board

Evaluation

- Head of Data and KIMS teams will inform the Information Governance Board of the new request and provide the request document
- Information Governance Board will evaluate the request document and verify whether scraping can take place

Web scraping log

Head of Data and KIMS teams will update the central scraping log with the approved activities and inform the business unit they can begin scraping. This is to ensure that the correct level of due dilligence is assigned to each project and that all activities are recorded for auditability and accountability purposes in the Web Scraping Log.

Web scraping

Web scraping activities can commence and will only be undertaken in accordance with the Web Scraping Policy and corresponding best practices.

Points of contact

Knowledge and Information Management and Security

Information Commissioners Office

Glossary

Application Programming Interface (API)

In terms of web scraping, an API can be built by the owner of a website to allow easy access to data held on the site without the need for building a web scraper.

Depth of Crawl

Depth of crawl refers to how far into a websites page hiearchy a web-crawler accesses. A websites homepage is the top of this hierarchy (level 0), pages linked from the home page are at level 1, pages linked from level 1 pages are level 2 and so forth. Limiting the depth of crawl therefore means limiting which level the web-crawler will access.

Hypertext Markup Language

The standard markup language for defining the structure of web pages. HTML elements are used to tell the browser how to display the contents of a web page.

Idle Time

Idle time, or sleep time, is the pause between each request made to a website by the scraper.

Parallelising the Crawling

A parallel crawler is a crawler that runs multiple processes in parallel with the goal to maximise download rates and minimise scraping overheads. For example, if website A and B contain multiple webages, then a parallelised crawl might capture a page from website A, followed by a page from website B and so on. If crawls are not parallelised, then the overhead on a single website is likely to be vastly increased.

Robots Exclusion Protocol

Also known as the Robots.txt file. This is a standard used by websites to communicate with web crawlers and other web robots. Essentially it provides website owners a method to restrict part of their websites from scraping or limit scraping entirely to just certified scrapers such as search engines. More information on this subject can be found on the robots.txt website.

User Agent String

A User Agent String is a custom piece of text that can be added to a scraper allowing the website owner to identify the operator and/or the purpose of the web scraping activity being undertaken. This can be modified when creating the web scraper and is often used for transparencies sake.

Web Crawler

A web crawler, sometimes referred to as a spider, is an internet bot that systematically browses the World Wide Web for the purposes of information gathering. The most common example of this is search engines like Google who use web crawlers to keep their search results up-to-date.

Web Scraping

Web scraping is defined as the automatic collection of data retrieved from the internet using a piece of software or program.