BRUSSELS PRIVACY HUB
Vrije Universiteit Brussel

Web scraping: legal grounds

By Liubomir Nikiforov*

* PhD researcher in Law, Vrije Universiteit Brussel, lyubomir.nikiforov@vub.be

Contents

Abstract:

This contribution examines the legal grounds for web scraping, a technique involved in the training of AI models. First, a distinction between web scraping and web crawling is established in order to explore further the potential legal grounds applicable. The analysis identifies significant challenges in obtaining consent and questions the validity of legitimate interest as an alternative compliant ground for data scraping. Ultimately, this contribution concludes that both legal grounds pose significant challenges before scrapes.

1 Introduction

The use of Artificial Intelligence technology has been at the center of all debates recently. Undoubtedly, it has profound impact on many domains and this is the reason why governments, private entities and individuals put efforts in order to harness the potential stemming from this technology. In this contribution, I do not analyze the AI Act, the applications and implications of AI, but I explore the legal grounds for web scarping. This is a relevant topic because web scraping is a method employed to assemble the dataset, used for AI models' training. For this reason, I explain briefly what web scraping is in order to explore the potential legal grounds applicable.

2 Web scraping

Web scraping is used to extract specific data from web pages. It focuses on retrieving particular pieces of information from targeted web pages. This process involves accessing a set of webpages, examining their code, and extracting the desired data according to predefined patterns or rules. Finally, the data collected is organized in a structured format such as a database for a specific objective[1] such as price comparison or market research. The more targeted approach in finding information is what distinguishes web scraping from web crawling.

The latter involves a systematic search and collection of data. This process is typically carried out by browsing engines to create a comprehensive index of the web for search purposes.

[1] *Web scraping ed intelligenza artificiale generativa—Nota informativa e possibili azioni di contrasto* (pp. 1–7). (2024). Garante per la protezione dei dati personali. https://www.gpdp.it/home/docweb/-/docweb-display/docweb/10019984

The whole crawling technique could be simplified in the following way. First, the process begins with a list of websites, the *crawler* visits these websites and collect links to other webpages, which in turn visits and collects other links thereof. This information is stored in a database for the purposes of creating a thorough web of all the websites found. This is why the program used, and previously called, crawler, is also known as "spider".

Therefore, web crawling is the technique used to index web pages on searching engines, while web scraping is narrower and focuses on collecting specific data such as price or location. In reality, web crawling and web scraping can be combined to provide a comprehensive and efficient data collection. While crawling discovers and indexes relevant web pages, scraping extracts specific data from these pages. This distinction is technical and in the following lines, I refer only to web scraping.

There is nothing inherently reproachable in fetching information from web sites for commercial purposes as long as it is lawful. Those software tools are used in order to create a relevant training dataset for AI models (the most relevant of them, LLM). Datasets are usually created in-house or acquired. Thus, developers can use datasets obtained from their customers and/or through their own scraping activities. In addition, they can use data form third parties such as open repositories like Common Crawl, Hugging Face, or LAION AI.[2] Using the EDPB Taskforce differentiation, the following stages of training could be outlined: collection of training data, pre-processing of the data, training, prompts and output, and finally training with prompts.[3]

When it comes to focus of this contribution, the collection of personal data, which may be privacy or commercially sensitive, automatic data retrieval for AI datasets training purposes, seems to raise concerns related to the specific legal grounds it relies on, and thus lawfulness of the processing. This is the main point of this contribution.

[2] *Web scraping ed intelligenza artificiale generativa—Nota informativa e possibili azioni di contrasto* (pp. 1–7). (2024). Garante per la protezione dei dati personali. https://www.gpdp.it/home/docweb/-/docweb-display/docweb/10019984

[3] *Report of the work undertaken by the ChatGPT Taskforce 23 May 2024* (pp. 1–14). (2024). European Data Protection Board (EDPB). https://www.edpb.europa.eu/our-work-tools/our-documents/other-guidance/report-work-undertaken-chatgpt-taskforce_en

3 Legal grounds: consent or legitimate interest

To ensure a comprehensive understanding, it is important to clarify the territorial scope of the GDPR. According to Article 3 of the GDPR, the regulation applies to data processing activities of controllers and processors in the EU, irrespective of whether the processing takes place in the EU or not. This broad scope means that web scraping activities targeting EU data subjects fall under GDPR, even if conducted from outside the EU. For the sake of this contribution, I do not examine in detail the territorial scope of the GDPR for the purposes of web scraping. Conversely, I assume a context where an organization retrieves as much data as possible from the Internet and therefore, it is highly likely that it processes personal data from data subjects within the EU. In addition, I assume that the scraper is the data controller at least of part of the processing as well as that the software product is deployed within the EU market. This latter variable is another reason for the application of the GDPR. For all of that, in the EU, in almost every case the GDPR would be applicable given the above-mentioned variables.

However, does all this really matter given that people have made publicly available their data online and Art. 9(e)? It is a tempting to answer negatively. First, not long ago, the Court of Justice of the European Union decided that websites can use their terms and conditions agreement to prohibit scraping of their data.[4] Thus, already some service providers explicitly oppose their data being collected. The Italian data protection authority points out that from one side, the protection provided by locked-in profiles could be a potential guardrail in order to prevent excessive web scraping. From the other side, an overreliance on this method may lead to excessive data collection by service providers in violation of the principle of data minimisation in Art. 5 (1)(c).[5] Second, it should not be overlooked that users' data has been acquired for purposes other than scraping, which mandates that the scraper indicates specific legal grounds for its activities. Third, scraping is a separate data processing activity, which clearly has an economic dimension. Forth, it is highly debatable whether users could have reasonable expectations that their data may be collected for scraping activities when

[4] Judgement of 15 January 2015, Ryanair Ltd v PR Aviation BV, Case C-30/14, EU:C:2015:10. https://curia.europa.eu/juris/document/document.jsf?text=&docid=161388&pageIndex=0&doclang=EN&mode=req&dir=&occ=first&part=1&cid=340454

[5] *Web scraping ed intelligenza artificiale generativa—Nota informativa e possibili azioni di contrasto* (pp. 1–7). (2024). Garante per la protezione dei dati personali. https://www.gpdp.it/home/docweb/-/docweb-display/docweb/10019984

uploading their personal data online. Moreover, there are multiple cases where available data in the Internet may not be considered publicly available. For example, it is plausible that someone else than the one to whom the data refers may have shared this data or in the cases where information is locked behind a private account.[6] All of this leads to the conclusion that publicly available personal data could not be used freely, or this is to say without proper assessment of the context[7] and the legal grounds applicable therefor.

Recently the Dutch data protection authority (DPA) has issued guidelines concerning the legal grounds for data processing when it comes to scraping.[8] As the main legal grounds discussed in literature as well as in the guidelines are legitimate interest (Art. 6(f), GDPR) and consent (Art. 6(a), GDPR), the following lines will be dedicated to their exploration.

3.1 Legitimate interest

The Dutch DPA's guidelines summarize the requirements for an adequate use of legitimate interest as a legal basis, Art. 6(1)(f) of the GDPR. Those conditions should be met cumulatively and are the following:

1. There is a legitimate interest of the scraping party or a third party.
2. The processing is necessary for the pursuit of this legitimate interest.
3. The interests or fundamental rights and freedoms of the data subjects do not override your legitimate interest or that of the third party.

Concerning the first point, there is no exclusive list of what a legitimate interest is. The GDPR provides a very limited clarification in Recitals 47-49. However, there are some authoritative

[6] *Handreiking Scraping door particulieren en private organisaties*. (2024). Autoriteit Persoonsgegevens. https://www.autoriteitpersoonsgegevens.nl/actueel/ap-scraping-bijna-altijd-illegaal; https://www.autoriteitpersoonsgegevens.nl/system/files?file=2024-05/Handreiking%20scraping%20door%20particulieren%20en%20private%20organisaties.pdf

[7] *Report of the work undertaken by the ChatGPT Taskforce 23 May 2024* (pp. 1–14). (2024). European Data Protection Board (EDPB). https://www.edpb.europa.eu/our-work-tools/our-documents/other-guidance/report-work-undertaken-chatgpt-taskforce_en

[8] *Handreiking Scraping door particulieren en private organisaties*. (2024). Autoriteit Persoonsgegevens. https://www.autoriteitpersoonsgegevens.nl/actueel/ap-scraping-bijna-altijd-illegaal; https://www.autoriteitpersoonsgegevens.nl/system/files?file=2024-05/Handreiking%20scraping%20door%20particulieren%20en%20private%20organisaties.pdf

guidelines, which could be used. For example, Opinion 6/2014 of the Article 29 Working Party[9] provides some directions:

1. exercise of the right to freedom of expression or information, including in the media and the arts
2. conventional direct marketing and other forms of marketing or advertisement
3. unsolicited non-commercial messages, including for political campaigns or charitable fundraising
4. enforcement of legal claims including debt collection via out-of-court procedures
5. prevention of fraud, misuse of services, or money laundering
6. employee monitoring for safety or management purposes
7. whistle-blowing schemes
8. physical security, IT and network security
9. processing for historical, scientific or statistical purposes
10. processing for research purposes (including marketing research)

The Dutch DPA provides a similar list[10], based on the Charter of Fundamental Rights of the European Union. However, in the guidelines concerning scraping and legal grounds it is highlighted that commercial interest in processing personal data does not match a valid legitimate interest. This means that an interest other than marketing or communication with a client will not qualify as a legitimate interest.

It is clear that particular exceptions may be envisaged, however, when it comes to web scraping of personal data for the sake of creating a commercially valuable dataset, for example in order to create an AI model or to sell it to a private party aiming to employ it that way, Art. 6(1)(f) is not a valid legal ground.

As far as the other two conditions are concerned, necessity and other parties overriding rights, they would not apply as the first one is not completed, and for the sake of this contribution, they are not discussed.

[9] *Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC*. (2014). [Opinion]. WP29. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf

[10] https://www.autoriteitpersoonsgegevens.nl/themas/basis-avg/avg-algemeen/grondslagen-avg-uitgelegd#grondslag-gerechtvaardigd-belang

This does not mean that the use of use of Art. 6(1)(f) of the GDPR is automatically discarded. The EDPB recalls that the first stages of web scraping are particularly risky in terms of fundamental rights and freedoms, and thus a thorough balancing test should be carried out.[11] According to the EDPB provided that certain safeguards such as "defining precise collection criteria and ensuring that certain data categories are not collected or that certain sources (such as public social media profiles) are excluded from data collection", as well as deletion and anonymization of personal data before the training stage. Although those safeguards could be a subject to a further debate of another contribution, the usefulness of legitimate interest as a legal ground for commercial web scraping is limited.

3.2 Consent

Consent (Art. 6(1)(a), GDPR) is the other potential option for a valid legal basis. The main challenge, however, is to prove the validity of consent. This is so for two reasons. From one side, challenging to identify each person whose data has been collected. This not only involves high transaction costs, but also may be virtually impossible. From the other side, even if complete identification of every data subject, the elements constituting a valid consent pose specific hurdles. In the hypothetical situation where obtaining consent is possible from the data subject, the following challenges remain.

First, consent should be "freely given". Although it is challenging to frame this element in the current contribution due to the high level of abstraction, it is not difficult to imagine a context where user's consent would not be considered free. For example, all cases where scraped personal data has been collected through a bundled consent tied to the access of the service providers platform, would not qualify as a free consent. To illustrate, imagine a popular social media platform requiring users to consent to web scraping of their data to continue using the service. This is not considered freely given consent because users are coerced into agreeing to avoid losing access to the platform. True consent must allow users to refuse without facing negative consequences. This issue may be exacerbated by platforms with a dominant position or a lack of adequate alternative of the same service.

[11] *Report of the work undertaken by the ChatGPT Taskforce 23 May 2024* (pp. 1–14). (2024). European Data Protection Board (EDPB). https://www.edpb.europa.eu/our-work-tools/our-documents/other-guidance/report-work-undertaken-chatgpt-taskforce_en

Second, any consent request should be specific enough concerning the purposes pursued. This is already very high bar for many private entities using automated tool, especially in the domain of advertising. When it comes to web-scraping where personal data is collected for building a dataset whose objective may not be known from the start, reinforces this issue.

Third, users should be able to take an informed decision when providing consent. This is probably the most challenging element because it is highly plausible that data subjects may lack sufficient understanding of the information provided or the implications their approval entails.

Finally, consent should be an unambiguous expression of a person's will. As it was already mentioned above, the mere circumstance that a set of personal data is online does not constitute a manifestly publicly available personal data. Thus, scrapers need to find be able to prove that a user willingly, actively and knowingly allowed the collection of the data. There is no single solution here as the experience with consent after the adoption of the GDPR proved. Despite that a click on a pop-up window usually constitutes a valid expression of a person's conscious choice, this hypothesis could be easily refuted based on the existence of nudging techniques, bundled consent forms, information overload, users' specific characteristics such as if he is a child.

In spite of those barriers, it is possible to obtain valid consent from users. However, I have started this section stating that in order to acquire consent, a web scraper would need to identify all individuals whose data is targeted. This is possible only in this contribution as an abstract examination, but in practice, it is virtually impossible to carry out due to the prohibitive transaction costs it would suppose, and hence, the exception in Art. 14(5)(b) would apply.

Therefore, given the framework of this contribution, namely, a web-scraping entity which is a data controller and which pursues economic benefit out of its activities would not be able to rely on Art. 6(1)(a) and (f) of the GDPR.

4 Conclusion

This contribution explores the role of web scraping as a method for feeding training data for AI models, and the associated legal challenges under GDPR. The examination reveals that

obtaining valid consent poses difficulties such as proper identification of all data subjects and ensuring informed consent. Additionally, the reliance on legitimate interest as a legal basis is limited, particularly for commercial purposes. Consequently, this contribution concludes that third party scraping commercial practices do not satisfy completely the requirements of the two legal grounds discussed.