Ethical & Legal checklist for web scraping

— BEFORE

Do I need to scrape this data?

Have I checked that there isn't an API I can use instead?

Can I get the data from a press officer or data portal?

Are there terms of use? If so, have I read them?

Do they expressly prohibit scraping? If not, what do they allow / not allow?

If I'm ignoring the terms of service, is there a public interest justification?

Is there anything in the robots.txt file I should be aware of? (Overview of robots.txt here)

Am I infringing copyright by scraping any of the material?

Cardiff University for example says on its website — under Terms of Use — that:
"All text, images and other content on this website is copyright of Cardiff University unless explicitly stated otherwise. It may only be downloaded or copied without first obtaining permission for the purposes of teaching, administration and research within the University, or for personal, non-commercial use. If you wish to reproduce our website content in any other way, or for any other purpose, you must first contact the Digital Communications team for permission."

Data protection

If there is personal data involved, am I working on the basis that the people have not consented to third-party (that's me in this case) use of this data? Or can I continue on this basis that this is now public information?

If it is personal data, how am I storing it securely?

Are there any other aspects to DPA / GDPR that I need to take into account?

Development

Am I using a cache when rerunning my calls to the website or storing the returned material in a variable? (See the requests_cache library)

— DURING

Running

Am I running my scraper in peak hours or in off-peak hours?

Am I overwhelming or pressurising the server by overly frequent requests? Have I designed pauses between calls?

Have I set a timeout on my requests?

Have I fixed the number of concurrent requests per domain?

Do I need to use auto throttling to slow my requests during busy periods (scrapy)?

Identification

Have I identified myself in the headers? Do I need to? Check your own user agent with a site like: https://www.whatismybrowser.com/detect/what-is-my-user-agent

Am I using User-Agent rotation - if so, is it editorially justified?

— AFTER

Will there be problems if I publish the data?

If I publish the data, have I got sufficient documentation to go with it?

Will there be problems if I publish the scraper or the scraping code?

Will I need to rerun my scraper for updated data?

— NOTES

See some practical examples here of these points in code:
https://colab.research.google.com/drive/1qplgB2-g8dwYjO5O6zzRZio7wFrc8LXU?usp=sharing

Latest guidance from the Associated Press is:

Sometimes, a data provider's website allows users to browse or search a data set but fails to provide direct download of the data. In this situation it may be possible to use software to step through the pages of the site and extract the data in a process known as web scraping.

Some website operators sanction this practice, and others oppose it. A website with policies limiting or prohibiting scraping often will include them in its terms of service or in a 'robots.txt' file, and reporters should take these into account when considering whether to scrape.

Scraping a website can cause its servers to work unusually hard, and in extreme cases, scraping can cause a website to stop working altogether and treat the attempt as a hostile attack. Therefore, follow these precautions:

— Scraping should be seen as a last resort. First try to acquire the desired data by request it directly.
— Limit the rate at which the scraper software requests pages in order to avoid causing undue strain on the website's servers.
— Wherever feasible identify yourself to the site's maintainers by adding yur contact information to the scraper's requests via the HTTP headers.

— *Associated Press Stylebook*, 55th edition, 2020-22, p. 358-39