zyte

ZYTEPAPER

Web scraping: Best practices

Ensuring your web scraping project stays out of trouble.

Introduction

At Zyte, we care about ensuring that our services respect the rights of websites and companies whose data we scrape. We hear a lot that scraping is a legal grey area, but the truth is scraping itself isn't illegal. It's the manner in which you scrape and what you scrape that falls into the grey area.

In this guide, we'll give you a set of guidelines to follow when scraping the web so you know when you need to be cautious about the manner and type of data you scrape.

Disclaimer: We are not your lawyer, and the recommendations in this guide do not constitute legal advice. Our Head of Legal is a lawyer, but she's not your lawyer, so none of her opinions or recommendations in this guide constitute legal advice from her to you.

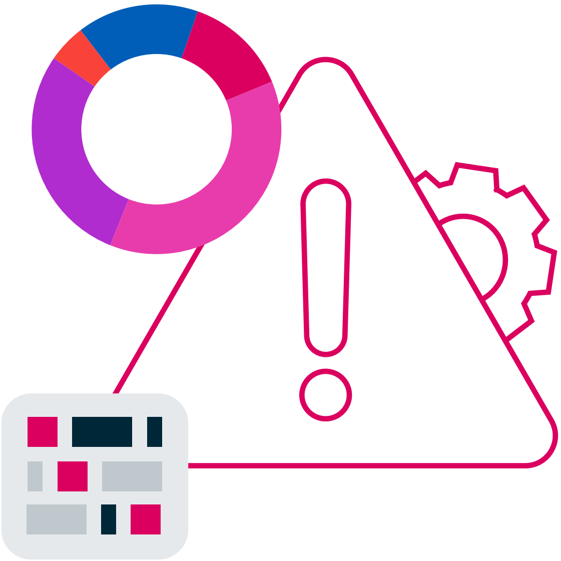
The commentary and recommendations outlined below are based on Zyte's experience helping our clients (startups to Fortune 100's) maintain legal compliance whilst scraping 7 billion web pages per month. If you want assistance with your specific situation then you should consult a lawyer.

Best practice #1

Don't be a burden

The first rule of scraping the web is **do not harm the website**. The second rule of web crawling is **do NOT harm the website.**

This means that the volume and frequency of queries you make should not burden the website's servers or interfere with the website's normal operations.



You can accomplish this in a number of ways:

Limit the number of concurrent requests to the same website from a single IP.

Respect the delay that crawlers should wait between requests by following the crawl-delay directive outlined in the robots.txt file.

If possible it is more respectful if you can schedule your crawls to take place at the websites off-peak hours.

A crucial aspect of this rule is providing the web administrators of the websites you scrape an easy way to contact you. At Zyte we accomplish this via making an abuse report available on our website.

If you ever receive an abuse report from a website you are scraping you should either stop scraping the site or limit the scraping in order to rectify the abuse reported.

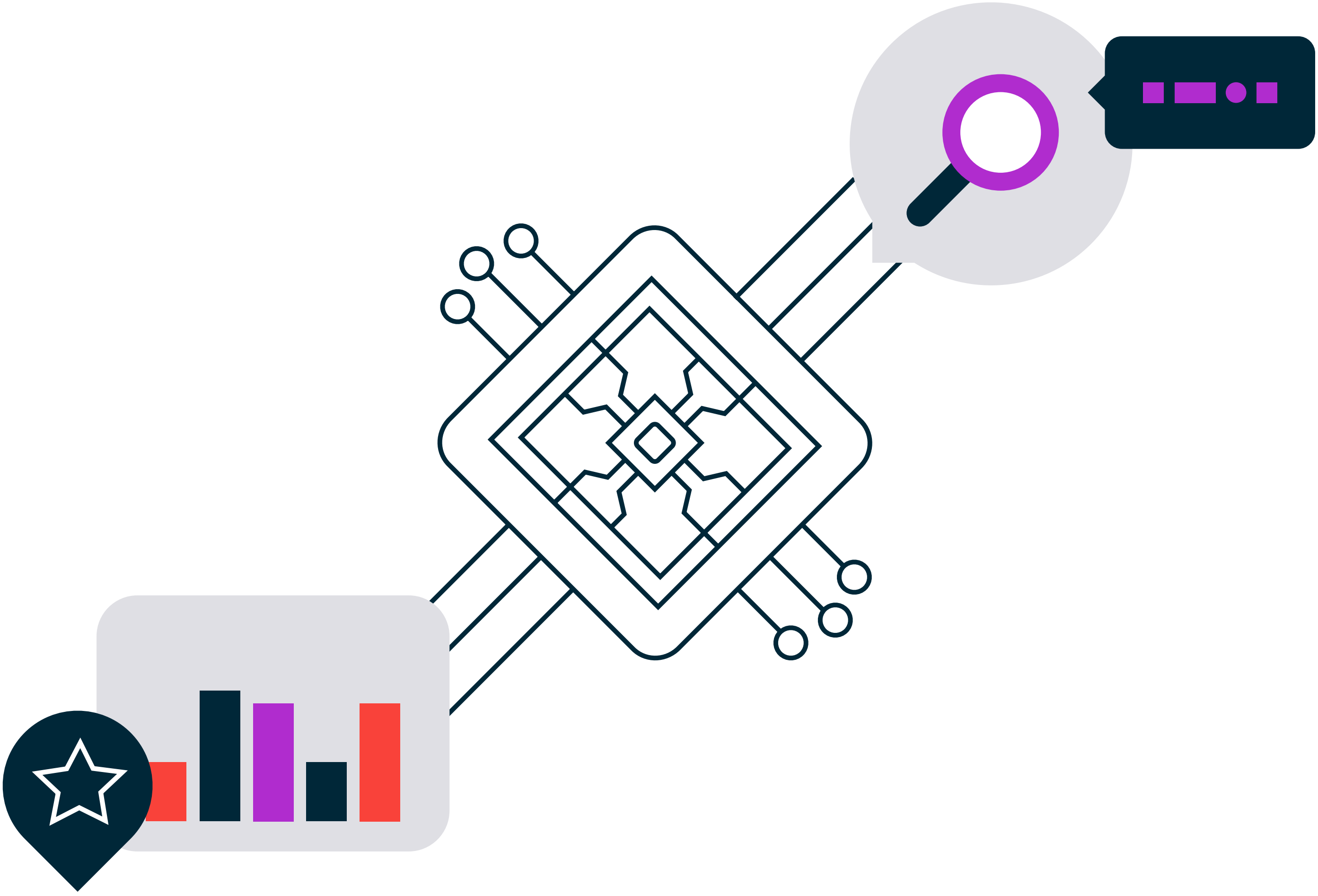
Best practice #2

Always inspect the robots.txt

When planning a web scraping project, your first step should **always check the robots.txt file** (usually available at the root of a website - www.example.com/robots.txt).

This document describes what a crawler should or shouldn't crawl according to the Robots Exclusion Standard.

The robots file also specifies what is considered as good behaviour on that site, such as areas that are allowed to be crawled, restricted pages, and frequency limits for crawling.



Best practice #3

Don’t violate copyright

When scraping a website you should always consider whether the web data you are planning to extract is copyrighted.

Copyright is defined as the exclusive legal right over a physical piece of work — like an article, picture, movie, etc. It basically means, if you create it, you own it. In order to be copyrightable, the work needs to be original and tangible.

The common types of material on the web that might be copyrighted are: articles, videos, pictures, stories, music, databases.

As a result, copyright is very relevant to scraping because much of the data on the internet (like articles and videos) are copyrighted works.

However, there are some situations when exceptions can apply to all or part of the data enabling it to be legally scraped without infringing on the owner’s copyright:

Fair Use

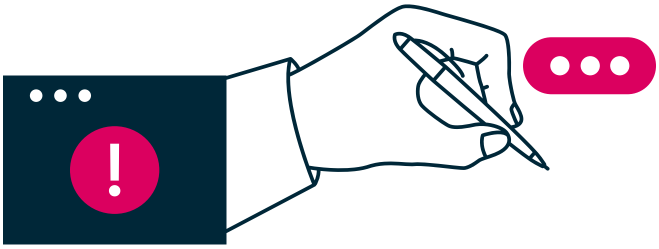
Fair Use is an exception that permits limited use of copyrighted material. Typically, fair use includes categories such as criticism/parody, comment, news reporting, teaching, scholarship, and research. One example of fair use is the publishing of short snippets of articles with links, which is generally okay under the fair use exception due to the transformative and limited nature of the use.

Transformative Use

One factor in determining fair use is whether the usage is transformative. Instead of distributing and storing exact duplicates or lengthy portions of the crawled website, transform the content and the use of the content in some way so that you are not violating copyright.

Facts

The facts within copyrighted material are often not covered by copyright laws, so if you limit what is being scraped to just the factual matters -- ie names of products, price, etc, then it is acceptable to scrape.



Note: different countries have different exceptions to copyright law, and you should always ensure that an exception applies within the jurisdiction within which you’re operating.

Best practice #4

Don’t breach GDPR

The introduction of GDPR completely changes how you can scrape the personal data of EU citizens (and sometime non-EU citizens as well). For a deeper explanation of how GDPR affects web scrapers, be sure to check out our Web Scrapers Guide to GDPR.

However, in this section we will briefly outline the best practices when it comes to scraping personal data. Personal data is any data that can identify an individual person: Name, Email, Phone Number, Address, User Name, IP Address, Bank or Credit Card Info, Medical Data, Biometric Data.

Unless, you have a “lawful reason” to scrape and store this data you will be in breach of GDPR if any of the scraped data belongs to EU residents. In the case of web scraping, the most common legal reasons are **legitimate interest** and **consent**.

Consent

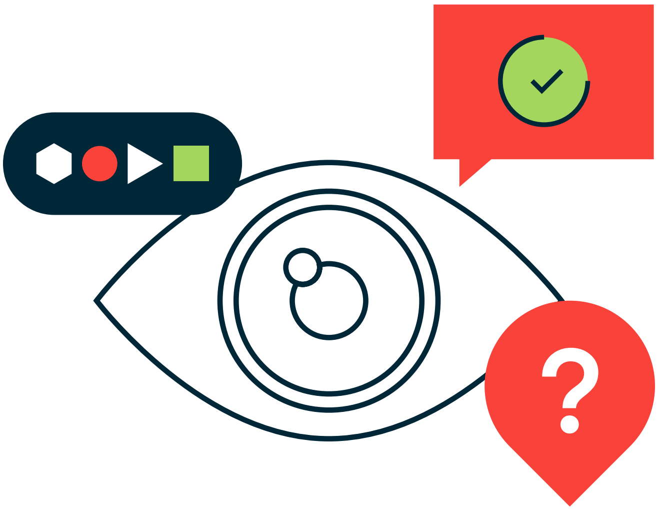
For consent to be your lawful reason to scrape a person’s data, you need to have that person’s explicit consent to scrape, store and use their data in the way you intended. This means that you or a 3rd party must have been in direct contact with the person and they agreed to terms that allow you to scrape their data.

An example of this would be companies like Mint.com, where users give Mint consent to log into their online banking accounts and retrieve their banking transactions so that they can be tracked and displayed in a more userfriendly format on Mint.com.

Legitimate Interest

For most companies, it will be very difficult for you to demonstrate that you have a legitimate interest in scraping someone’s personal data.

In most cases, only governments, law enforcement agencies, etc. will have what would be deemed to be a legitimate interest in scraping the personal data of its citizens as they will typically be scraping people’s personal data for the public good.



Best practice #5

Show yourself

It is always best practice to identify yourself whenever possible and put contact details in the crawler's header. It's important that when using data center or other third party IPs, ensure that they can get an abuse report or cease and desist back to you if needed. If you don't, they'll have to dig into their logs and look for the offending IPs.

Be nice to the friendly sysadmins in your life and identify your crawler. When using Scrapy this can be accomplished using the USER_AGENT setting. Here you can share your crawler name, company name and a contact email:

```
USER_AGENT = 'MyCompany-MyCrawler (bot@mycompany.com)'
```

Then on your website you should provide an easy to use contact form or abuse report where a sysadmin can let you know about their issue with your web scraping.

Best practice #6

Beware of login and website T&CS

When you login and/or explicitly agree to a website's terms and conditions you are entering into a contract with the website owner, thereby agreeing to their rules regarding web scraping. Which can explicitly state that you aren't allowed to scrape any data on the website.

This means that you need to carefully review the terms and conditions you are agreeing to if your spiders have to login to scrape data, as they could stipulate that you're not allowed to scrape their data.

You should always honor the terms of any contract you enter into, including website terms and conditions and privacy policies.

At Zyte we turn websites into data with industry leading technology and services.

Our solutions include:

- **Data Extraction Service**
  Let our web scraping experts build and manage the bespoke data extraction solution for your business needs.
- **Automatic Extraction powered by AI**
  Instantly access accurate web data through our user-friendly interface or various Extraction APIs and save time getting the data you need.
- **Smart Proxy Manager (formerly Crawlera)**
  Forget about proxy lists. We manage hundreds of thousands of proxies, so you don't have to.
- **Data extraction platform**
  Access developer tools, data extraction APIs and documentation, built and maintained by our world-leading team of over 100 extraction experts.

zyte

It's yours. The web data you need.

Access clean, valuable data with web scraping services that drive your business forward.

Talk to us

www.zyte.com

Cuil Greine House, Ballincollig Commercial
Park Link Road, Ballincollig , Co. Cork, Ireland