MENU

Web Scraping and Crawling Are Perfectly Legal, Right?

18 APRIL 2017 on scraping, crawling, legal, law, lawsuit, tos, harvesting, data

"Come on, I worked so hard on this project! And this is publicly accessible data! There's certainly a way around this, right? Or else, I did all of this for nothing... Sigh..."

Yep - this is what I said to myself, just after realizing that my ambitious data analysis project could get me into hot water. I intended to deploy a large-scale web crawler to collect data from multiple high profile websites. And then I was planning to publish the results of my analysis for the benefit of everybody. Pretty noble, right? Yes, but also pretty risky.

Interestingly, I've been seeing more and more projects like mine lately. And even more tutorials encouraging some form of web scraping or crawling. But what troubles me is the appalling widespread ignorance on the legal aspect of it.

So this is what this post is all about - understanding the possible consequences of web scraping and crawling. Hopefully, this will help you to avoid any potential problem.

Disclaimer: I'm not a lawyer. I'm simply a programmer who happens to be interested in this topic. You should seek out appropriate professional advice regarding your specific situation.

What are web scraping and crawling?

Let's first define these terms to make sure that we're on the same page.

1. **Web scraping**: the act of automatically downloading a web page's data and extracting very specific information from it. The extracted information can be stored pretty much anywhere (database, file, etc.).
2. **Web crawling**: the act of automatically downloading a web page's data, extracting the hyperlinks it contains and following them. The downloaded data is generally stored in an index or a database to make it easily searchable.

For example, you may use a web scraper to extract weather forecast data from the National Weather Service. This would allow you to further analyze it.

In contrast, you may use a web crawler to download data from a broad range of websites and build a search engine. Maybe you've already heard of Googlebot, Google's own web crawler.

So web scrapers and crawlers are generally used for entirely different purposes.

Why is web scraping often seen

negatively?

The reputation of web scraping has gotten a lot worse in the past few years, and for good reasons:

1. It's increasingly being used for business purposes to gain a competitive advantage. So there's often a financial motive behind it.
2. It's often done in complete disregard of copyright laws and of Terms of Service (ToS).
3. It's often done in abusive manners. For example, web scrapers might send much more requests per second than what a human would do, thus causing an unexpected load on websites. They might also choose to stay anonymous and not identify themselves. Finally, they might also perform prohibited operations on websites, like circumventing the security measures that are put in place to automatically download data, which would otherwise be inaccessible.

Tons of individuals and companies are running their own web scrapers right now. So much that this has been causing headaches for companies whose websites are scraped, like social networks (e.g. Facebook, LinkedIn, etc.) and online stores (e.g. Amazon). This is probably why Facebook has separate terms for automated data collection.

In contrast, web crawling has historically been used by the well-known search engines (e.g. Google, Bing, etc.) to download and index the web. These companies have built a good reputation over the years, because they've built indispensable tools that add value to the websites they crawl. So web crawling is generally seen more favorably, although it may sometimes be used in abusive ways as well.

So is it legal or illegal?

Web scraping and crawling aren't illegal by themselves. After all, you could scrape or crawl your own website, without a hitch.

The problem arises when you scrape or crawl the website of somebody else, without obtaining their prior written permission, or in disregard of their Terms of Service (ToS). You're essentially putting yourself in a vulnerable position.

Just think about it; you're using the bandwidth of somebody else, and you're freely retrieving and using their data. It's reasonable to think that they might not like it, because what you're doing might hurt them in some way. So depending on many factors (and what mood they're in), they're perfectly free to pursue legal action against you.

I know what you may be thinking. "Come on! This is ridiculous! Why would they

sue me?". Sure, they might just ignore you. Or they might simply use technical measures to block you. Or they might just send you a cease and desist letter. But technically, there's nothing that prevents them from suing you. This is the real problem.

Need proof? In Linkedin v. Doe Defendants, Linkedin is suing between 1-100 people who anonymously scraped their website. And for what reasons are they suing those people? Let's see:

1. Violation of the Computer Fraud and Abuse Act (CFAA).
2. Violation of California Penal Code.
3. Violation of the Digital Millennium Copyright Act (DMCA).
4. Breach of contract.
5. Trespass.
6. Misappropriation.

That lawsuit is pretty concerning, because it's really not clear what will happen to those "anonymous" people.

Consider that if you ever get sued, you can't simply dismiss it. You need to defend yourself, and prove that you did nothing wrong. This has nothing to do

with whether or not it's fair, or whether or not what you did is really illegal.

Another problem is that law isn't like anything you're probably used to. Because where you use logic, common sense and your technical expertise, they'll use legal jargon and some grey areas of law to prove that you did something wrong. This isn't a level playing field. And it certainly isn't a good situation to be in. So you'll need to get a lawyer, and this might cost you a lot of money.

Besides, based on the above lawsuit by LinkedIn, you can see that cases can undoubtedly become quite complex and very broad in scope, even though you "just scraped a website".

The typical counterarguments brought by people

I found that people generally try to defend their web scraping or crawling activities by downplaying their importance. And they do so typically by using the same arguments over and over again.

So let's review the most common ones:

1. *"I can do whatever I want with publicly accessible data."*

False. The problem is that the "creative arrangement" of data can be copyrighted, as described on cendi.gov:

> *Facts cannot be copyrighted. However, the creative selection, coordination and arrangement of information and materials forming a database or compilation may be protected by copyright. Note, however, that the copyright protection only extends to the creative aspect, not to the facts contained in the database or compilation.*

So a website - including its pages, design, layout and database - can be copyrighted, because it's considered as a creative work. And if you scrape that website to extract data from it, the simple fact of copying a web page in memory with your web scraper might be considered as a copyright violation.

In the United States, copyrighted work is protected by the Digital Millenium Copyright Act (DMCA).

2. *"This is fair use!"*

   This is a grey area:

   - In Kelly v. Arriba Soft Corp., the court found that the image search

engine Ditto.com made fair use of a professional photographer's pictures by displaying thumbnails of them.
    - In Associated Press v. Meltwater U.S. Holdings, Inc., the court found that Meltwater's news aggregator service didn't make fair use of Associated Press' articles, even though scraped articles were only displayed as excerpts of the originals.
3. *"It's the same as what my browser already does! Scraping a site is not technically different from using a web browser. I could gather data manually, anyway!"*

   False. Terms of Service (ToS) often contain clauses that prohibit crawling/scraping/harvesting and automated uses of their associated services. You're legally bound by those terms; it doesn't matter that you could get that data manually.

4. *"The worse that might happen if I break their Terms of Service is that I might get banned or blocked."*

   This is a grey area:

    - In Facebook v. Pete Warden, Facebook's attorney threatened Mr. Warden to sue him if he published his dataset comprised of hundreds of million of scraped Facebook profiles.

- In Linkedin Corporation v. Michael George Keating, Linkedin blocked Mr. Keating from accessing Linkedin because he had created a tool that they **thought** was made to scrape their website. They were wrong. But yet, he has never been able to restore his account. Fortunately, this case didn't go further.
- In LinkedIn Corporation v. Robocog Inc, Robocog Inc. (a.k.a. HiringSolved) was ordered to pay 40000$ to Linkedin for their unauthorized scraping of the site.

5. *"This is completely unfair! Google has been crawling/scraping the whole web since forever!"*

   True. But law has apparently nothing to do with fairness. It's based on rules, interpreted by people.

6. *"If I ever get sued, I'll Good-Will-Hunting my way into defending myself."*

   Good luck! Unless you know law and legal jargon extensively. Personally, I don't.

7. *"But I used an automated script, so I didn't enter into any contract with the website."*

This is a grey area:

- In Internet Archive v. Suzanne Shell, Internet Archive was found guilty of breach of contract while copying and archiving pages from Mrs. Shell's website using its web crawlers. On her website, Mrs. Shell displays a warning stating that as soon as you copy content from her website, you enter into a contract, and you owe her 5000$US per page copied (!!!). The two parties apparently reached an amicable resolution.
- In Southwest Airlines Co. v. BoardFirst, LLC, BoardFirst was found guilty of violating a browsewrap contract displayed on Southwest Airlines' website. BoardFirst had created a tool that automatically downloaded the boarding passes of Southwest's customers to offer them better seats.

8. *"Terms of Service (ToS) are not enforceable anyway. They have no legal value."*

False. The Bingham McCutchen LLP law firm published a pretty extensive article on this matter and they state that:

> *As is the general rule with any contract, a website's terms of use will generally be deemed enforceable if mutually agreed to by the parties. [...] Regardless of whether a website's terms of use are clickwrap or browsewrap, the defendant's failure to read those terms is generally found irrelevant to the enforceability of its terms. One court disregarded arguments that awareness of a website's terms of*

> *use could not be imputed to a party who accessed that website using a web crawling or scraping tool that is unable to detect, let alone agree, to such terms. Similarly, one court imputed knowledge of a website's terms of use to a defendant who had repeatedly accessed that website using such tools. Nevertheless, these cases are, again, intensely factually driven, and courts have also declined to enforce terms of use where a plaintiff has failed to sufficiently establish that the defendant knew or should have known of those terms (e.g., because the terms are inconspicuous), even where the defendant repeatedly accessed a website using web crawling and scraping tools.*

In other words, Terms of Service (ToS) will be legally enforced depending on the court, and if there's sufficient proof that you were aware of them.

9. *"I respected their robots.txt and I crawled at a reasonable speed, so I can't possibly get into trouble, right?"*

   This is a grey area.

   robots.txt is recognized as a "technological tool to deter unwanted crawling or scraping". But whether or not you respect it, you're still bound to the Terms of Service (ToS).

10. *"Okay, but this is for personal use. For my personal research only. I won't re-publish it, or publish any derivative dataset, or even sell it. So I'm good to go, right?"*

    This is a grey area. Terms of Service (ToS) often prohibit automatic data collection, for any purpose.

    According to the Bingham McCutchen LLP law firm:

> *The terms of use for websites frequently include clauses prohibiting access or use of the website by web crawlers, scrapers or other robots, including for purposes of data collection. Courts have recognized causes of action for breaches of contract based on the use of web crawling or scraping tools in violation of such provisions.*

11. *"But the website has no robots.txt. So I can do what I want, right?"*

    False. You're still bound to the Terms of Service (ToS), and the content is copyrighted.

General advice for your scraping or crawling projects

Based on the above, you can certainly guess that you should be extra cautious with web scraping and crawling.

Here are a few pieces of advice:

1. Use an API if one is provided, instead of scraping data.
2. Respect the Terms of Service (ToS).
3. Respect the rules of *robots.txt*.
4. Use a reasonable crawl rate, i.e. don't bombard the site with requests. Respect the *crawl-delay* setting provided in *robots.txt*; if there's none, use a conservative crawl rate (e.g. 1 request per 10-15 seconds).
5. Identify your web scraper or crawler with a legitimate user agent string. Create a page that explains what you're doing and why, and link back to the page in your user agent string (e.g. 'MY-BOT (+https://yoursite.com/mybot.html)')
6. If ToS or *robots.txt* prevent you from crawling or scraping, ask a written permission to the owner of the site, **prior** to doing anything else.
7. Don't republish your crawled or scraped data or any derivative dataset without verifying the license of the data, or without obtaining a written permission from the copyright holder.
8. If you doubt on the legality of what you're doing, don't do it. Or seek the

advice of a lawyer.

9. Don't base your whole business on data scraping. The website(s) that you scrape may eventually block you, just like what happened in Craigslist Inc. v. 3Taps Inc..
10. Finally, you should be suspicious of any advice that you find on the internet (including mine), so please consult a lawyer.

Remember that companies and individuals are perfectly free to sue you, for whatever reasons they want. This is most likely not the first step that they'll take. But if you scrape/crawl their website without permission and you do something that they don't like, you definitely put yourself in a vulnerable position.

Conclusion

As we've seen in this post, web scraping and crawling aren't illegal by themselves. They might become problematic when you play on somebody else's turf, on your own terms, without obtaining their prior permission. The same is true in real life as well, when you think about it.

There are a lot of grey areas in law around this topic, so the outcome is pretty unpredictable. Before getting into trouble, make sure that what you're doing

respects the rules.

And finally, the relevant question isn't "Is this legal?". Instead, you should ask yourself "Am I doing something that might upset someone? And am I willing to take the (financial) risk of their response?".

So I hope that you appreciated my post! Feel free to leave a comment in the comment section below!

This post was featured on Hacker News, Reddit, Lobsters and in the Programming Digest newsletter. Thanks to everyone for your support and feedback!

128 Comments

Type Comment Here (at least 3 chars)

Name (optional) E-mail Website (optional)

Submit

Sean Lang · 7 years ago

> *Identify your web scraper or crawler with a legitimate user agent string. Create a page that explains what you're doing...*

If Suzanne Shell can inforce an absurd ToS "agreement" against a library like the Internet Archive, then letting these people know who you are is probably a bad idea. It seems like the only solution is to do your scraping anonymously or be a company as large as Google that can afford to deal with the legal system.

3 | **Reply**

Benoit Bernard · 7 years ago

If what you're doing is legitimate/legal, then what I said holds true. You should not be scared of announcing who you are and what you're doing. You would be surprised of how effective it can be to create a page as I suggested. This really clears out any confusion for everybody. I know that it might seem counterintuitive.

But if you know that what you're doing is suspicious/illegal, and you still want to proceed with it, then yes. What you're suggesting is probably the way to go. But I won't be the person responsible for suggesting it :)

| **Reply**

Sean Lang · 7 years ago

The Internet Archive is trying to preserve the history of the Internet for future generations to learn from... They even let you

exclude your site from the archive with a robots.txt entry. I can hardly imagine a more "legitimate" use for scraping, yet they were still able to be attacked with a court case. I don't think legitimacy has anything to do with it.

11 | **Reply**

Benoit Bernard · 7 years ago

> *I don't think legitimacy has anything to do with it.*

I agree - maybe "legitimacy" is not the right word. Let me rephrase what I've already said.

In most cases, announcing who you are and what you're doing will greatly decrease the chances of getting into trouble. This will show your "good faith". Otherwise, people don't know whether you're the "good guy" or the "bad guy".

> *I can hardly imagine a more "legitimate" use for scraping, yet they were still able to be attacked with a court case.*

I agree. The Internet Archive's mission is very noble, and super

important. Despite this, they got sued. It only reconfirms that you should be extra cautious, and ask a written permission whenever possible - at least for websites that explicitly prohibit you from crawling/scraping.

3 | **Reply**

Anonymous · 7 years ago

"If you doubt on the legality of what you're doing, don't do it.", more like get your ass behind a proxy and do it anyway. Its unfortunate that law is used to help broken systems survive.

2 | **Reply**

Benoit Bernard · 7 years ago

If you ever get caught of scraping behind a proxy, then you'll no longer be able to use the excuse "hey, I didn't know that your Terms of Service prevent the use of web scrapers! I didn't read them!".

I agree with the last part of what you said, though. Yep, it's unfortunate that law is so broken regarding technology. It hasn't adapted well. One very good example is patents.

1 | **Reply**

Sean Lang · 7 years ago

> *If you ever get caught of scraping behind a proxy, then you'll no longer be able to use the excuse "hey, I didn't know that your Terms of Service prevent the use of web scrapers!...*

Operating behind a proxy doesn't mean you read or didn't read the ToS. There are plenty of technical reasons to use a proxy... Anything from response compression / filtering, to providing encryption over untrusted networks, to ensuring redundant network access. Hiding your IP address is just one reason among many.

Of course, I don't think ignorance counts as an excuse if those cases you cited above are accurate, so you should probably just focus on not getting caught by someone who has the resources to drag you to court.

2 | **Reply**

Benoit Bernard · 7 years ago

Hey Sean - I like it when people feel involved in a topic. This makes things interesting. Thanks for your comments! :)

> *Operating behind a proxy doesn't mean you read or didn't read the ToS. There are plenty of technical reasons to use a proxy...*

Technically, you're right. But my point was that operating a web scraper behind a proxy might raise the doubts of people. They might question whether you have something to hide.

> *[...] you should probably just focus on not getting caught by someone who has the resources to drag you to court.*

After all I've read, I don't feel comfortable with this idea, personally. But ultimately, the decision belongs to you.

^ | v **Reply**

Anonymous · 6 years ago

Use a chain of proxies then. Pick countries politically opposed to each other, like russia to india to Brazil to Germany to the USA and slap in a few third world proxies just to be sure. South eastern Asia and northern Africa tend to be hard to pull records from.

-2 ^ | v **Reply**

Benoit Bernard · 6 years ago

Besides the fact that this would slow down your scraping by an order of magnitude, this would probably work. People would have a hard time identifying or locating you.

But, sure enough, some people will be motivated enough (read: have enough $$$) to track you down if they really want to (e.g. LinkedIn).

So, again, I would not advise anybody to do that.

 Reply

1 Hidden

Soufiane · 7 years ago

Consider that I'm myself the main maintainer of an open source project that helps to scrape google pages (google disallow people to scrape their indexes). I don't use this programe myself, I just provide other people the ability to do it. People that use my programe might be out of law.

The question that comes in my mind: Do you know if I am myself in an illegal position if people make a bad usage of my script?

 Reply

Benoit Bernard · 7 years ago

Hey Soufiane - thanks for stopping by :)

I can't give you advice regarding your specific situation, as I'm not a lawyer. However, based on my post, you can guess what might be the possible outcome(s):

- They might ignore you.
- They might cancel all of your Google accounts (including your Gmail), as indicated in their ToS.

> *We may suspend or stop providing our Services to you if you do not comply with our terms or policies or if we are investigating suspected misconduct.*

- They might go further, like sending you a cease-and-desist letter or other.

This is highly unpredictable.

 **Reply**

Anonymous · 7 years ago

Hi Benoit, Thanks for you answer, I will try to make further investigations on the question!

Reply

Brian · 7 years ago

I would think that creating a "tool" that enables "scraping" and other text capture functions, by itself is not illegal. What it's used for, is unknown. It's like the hammer manufacturer that sells a hammer that is used to damage something - it wasn't meant to be used for that purpose specifically - and the manufacturer can even specifically say "Do not use for damaging purposes". How could the builder of that "tool" be held responsible?

Reply

Benoit Bernard · 7 years ago

Hey Brian - you bring interesting points.

> *I would think that creating a "tool" that enables "scraping" and other text capture functions, by itself is not illegal.*

Actually, he said that he created a tool to scrape Google search results. This is definitely against Google's ToS. So he might get into trouble for

creating that tool and **making it available** to other people. As a reference, check out what happened in Linkedin Corporation v. Michael George Keating. If he had created a general purpose scraping tool, then things would probably be different.

But I realize that I misread Soufiane's question. He wasn't asking whether he could get into trouble for creating that tool. He was asking whether **he** could get into trouble if **other people** misbehaved with his tool. In theory, people are responsible for their own acts, so he might get away without any problem. But again, we can't be sure about that.

Reply

Brian · 7 years ago

Are gun manufacturers responsible for what users do with their products? Are cigarette manufacturers responsible for what happens to smokers? Some don't inhale, I suppose - they just want to look "cool". Also, tools that are able to scrape Google results usually aren't specifically for that purpose - they grab text from any type web page. I think "tool" makers need to be given the benefit of the doubt. Thanks for the discussion!

6 Reply

3 Hidden

Bob · 7 years ago

Great article and food for thought. I scrape a county website about 50 times on one occasion per year (the data changes daily for about a month). I looked at the TOS today and it's very short, about 3 bullet items designed to protect themselves without mention of data limitations. Their robots.txt lists about 20 directories, only one of which is currently valid.

 **Reply**

Benoit Bernard · 7 years ago

Hey Bob - thanks for your kind words!

A few pointers:

1. If nothing is mentioned concerning scraping/crawling, you can at least assume that the content of the page is copyrighted.
2. Checking out the license of the data would be a good idea, if there's one.
3. Asking for their permission would also be a good idea.
4. Some government datasets are distributed under an "open data"

license, so maybe you could download it directly, instead of scraping it.

I hope that it helps!

Reply

Poul · 7 years ago

What if you have a news site like the old digg.com where people submitted links to news, and you took the title and an excerpt and made a link to the original article ?

-1 Reply

Benoit Bernard · 7 years ago

This might be seen as fair use. However, some courts ruled that similar cases were not fair use (see Associated Press v. Meltwater U.S. Holdings, Inc.).

At the very least, creating a page that explains exactly what your scraper is doing (e.g. "This bot visited your page because somebody linked to your content on the 'XYZ' news aggregator."), and linking to it in your user agent string would be a good idea. But it doesn't provide any firm guarantee.

You should definitely consult a lawyer, especially if you plan to start a social news aggregator.

Disclaimer: I'm not a lawyer :)

Reply

José · 6 years ago

Hello, you article is great. What I understand is that Google can crawl and I don't see why others would not crawl too. APIs are a great tool to get data legally.

Also, you have to provide value in exchange of what you do with the data (that is a marketing idea).

I would like to know what you think about contacting a website you crawled an email from and have data about that you know you can help with ?

All the best,

José

Reply

Benoit Bernard · 6 years ago

Hey Jose - thanks for your comments.

> *What I understand is that Google can crawl and I don't see why others would not crawl too.*

Well, Google is:

1. A big, well-known and generally trusted company.
2. A company with deep pockets ($$$).
3. A company bringing value to the world, and they've been crawling the web since a long time ago.

This is mostly why people accept being crawled by them. Those 3 criteria probably don't apply to most other people/companies, though.

> *APIs are a great tool to get data legally.*

Yes, an API is a great alternative to crawling/scraping, given that one exists for the data that you need. But even with APIs, there are some legal hurdles. The data that you receive isn't copyrightable, but

arguably, the underlying database that it comes from **is** copyrighted. So in theory, somebody could sue you on the basis that you infringed their copyrighted database.

> *you have to provide value in exchange of what you do with the data (that is a marketing idea).*

This is a possible approach, but it's not 100% bulletproof. Sure, more people may welcome your crawling/scraping activities given that you bring them value. But consider that if you upset somebody, even if you provide value, they might decide to stop you. Wasn't Pete Warden trying to bring value to the world, including Facebook themselves, after all?

> *I would like to know what you think about contacting a website you crawled an email from and have data about that you know you can help with*

IMO, you're looking at it in the wrong order. Here's what I would do:

1. Determine if permission is needed to crawl/scrape.
2. If permission is indeed, get it in writing.
3. Crawl/scrape.

Reply

Anonymous · 6 years ago

I don't really agree with your points,

Because it's a big company it can afford to do whatever it wants?

I mean come on, either scraping is legal or it's not.

2 Reply

Benoit Bernard · 6 years ago

> *I don't really agree with your points.*

Fine. We can't all agree on the same things.

> *Because it's a big company it can afford to do whatever it wants?*

Yes, almost.

> *I mean come on, either scraping is legal or it's not.*

It's not as simple as that. Law isn't black and white; it's open to interpretation.

So if somebody doesn't like being crawled by you, they can have whatever recourse they want against you. This includes suing you, but it's most likely not the first thing that they'll do.

5 | **Reply**

Anonymous · 6 years ago

you are saying nonsense..It can do because it's a large company and because they did it first.. [...] Your logic is so wrong but you can't even see it. I ques you are the type of person who thinks the government should have total control over population because it can and people can trust them..Come on .. If he said he thinks you are wrong then you are wrong end of the story ..Even prove you are right or drop the discussion.You can't say:You are right we can't agree and say the same stuff..NLP much?

-8 | **Reply**

Benoit Bernard · 6 years ago

> *You are saying nonsense..It can do because it's a large company and because*

> *they did it first..*

Again, yes, big companies have been able to do pretty much what they wanted for the past two decades. That is, until somebody came after them.

Do you remember:

1. The antitrust lawsuit against Microsoft) in 2001?
2. The antitrust lawsuit against Google in 2015?
3. The recent scandal involving Facebook and Cambridge Analytica?
4. Etc.

I'm not endorsing these things - I'm just saying that the big guys have an advantage over everyone else. And since they've been doing their thing for a long time , people tend to "trust" them more than the little guy. And also, because they have a lot of money, they can get away with pretty much anything. This is unfortunate, but it's how the world works. That's basically my point.

> *Your logic is so wrong but you can't even see it.I ques you are the type of person who thinks the government should have total control over population*

> *because it can and people can trust them..Come on ..*

You're misinterpreting what I'm saying. And no, I'm not for a totalitarian government at all. Anyway, I'm not sure how this relates to the legality of web crawling and scraping.

> *If he said he thinks you are wrong then you are wrong end of the story ..Even prove you are right or drop the discussion. You can't say:You are right we can't agree and say the same stuff.*

Well, I guess that my entire post and my reply here both prove my point. In my mind, it's perfectly okay to disagree with somebody else; this is the basis of a discussion. And this is what we're having here - a discussion - right?

By the way, if you have suggestions about how I should improve my post, please tell me.

5 | **Reply**

3 Hidden

Pro user · 6 years ago

Thanks

Reply

Anonymous · 6 years ago

Very informative. Thanks. What do you think of Yummly? How did they get to copy recipes from so many websites?

Reply

Benoit Bernard · 6 years ago

Well, I can't say anything about Yummly since I don't know the specific method that they used to gather their recipes.

However, there are many startups out there who use web scrapers and crawlers without even knowing the legal implications of copying stuff. Things on the web are not as "free" as many people might think.

Reply

Ben Walters · 6 years ago

Great analysis! I've started a few projects of my own that involve scraping, only to encounter legal or technical limitations. It's disappointing, but understandable. It can be ultimately frustrating to try and use so many APIs

with terrible or non-existent documentation too - it's no wonder we resort to scraping it ourselves!

3 | **Reply**

Benoit Bernard · 6 years ago

Hey Ben - thanks for your feedback!

I feel that whenever an API is available, we should use it. However, as you said, APIs often come with their own limitations - a lacking documentation being one of them.

This is often why we resort to web scraping/crawling. But again, we often end up with other kinds of limitations.

I agree that it's frustrating. The web started as a free, open place. But this is arguably no longer the case; corporations now drive the whole thing, so you can't do whatever you want anymore. Being extra cautious is probably the way to go.

P.S.: I loved your blog post about your **catPI** project. Nice work! :)

1 | **Reply**

Ben Walters · 6 years ago

Thanks, Benoit! And thanks for checking out my blog! I really need to work on a version 2.0 of the cat feeder...

I was most recently looking at building a forum scraper to extract and display all the images from a given thread (extremely useful for lazy car enthusiasts like myself). But between the forum apparently "owning the content" and the lack of an API for the vBulletin software, I quickly abandoned the idea.

Another good example is this project recently posted to Reddit:

https://www.reddit.com/r/dataisbeautiful/comments/7jkb37/i_created_a_website_to_visualize_and_quantify/

The OP created an awesome visualizer for Amazon reviews, but most of the comments expressed concerns about scraping Amazon's content (as they are quick to point out THEY own the reviews on their site). I mean com'on!!

Reply

Ryan · 6 years ago

Great post, just seeing it now as I learn more about what I'm trying to do...

Here's a workflow I'm tinkering with - I'm curious if you think it plays well within the current legal framework:

Scripted: duckduckgo -> ducky a search string (their version of Google's "I'm feeling lucky" function) that navigates to a site, pulls the url (if it exists) and navigates to Google's cached content version to scrape.

At no point does it seem to me to violate ToS. Thoughts?

1 | **Reply**

Benoit Bernard · 6 years ago

Hi Ryan,

I don't know about Duckduckgo (DDG), but Google implicitly prohibits any form of scraping on their site. See their Terms of Service:

> *Don’t misuse our Services. For example, don’t interfere with our Services or*

> *try to access them using a method other than the interface and the instructions that we provide.*

Instead of scraping DDG or Google, maybe you could use something like http://commoncrawl.org/ ?

Reply

AQ · 6 years ago

Informative article!Thank you- Much appreciated-Data Science Student

Reply

Anonymous · 6 years ago

What does the recent injunction involving HiQ (permitted to continue to scrape Linkedin), as stated by a California Judge, have on performing the same type of scraping in Canada?

Reply

Benoit Bernard · 6 years ago

That story about HiQ vs Linkedin is quite interesting.

I tend to believe that the judge ruled in favor of HiQ because Linkedin

was simultaneously developing its own competing product **and** trying to block HiQ; the judge probably saw this as anti-competitive.

Now, considering that:

- Different judges take different decisions (law = rules interpreted by people).
- Laws are quite different from country to country.
- There aren't that many cases of jurisprudence about web crawling/scraping.
- Each case has its own context.

It would probably be a mistake to think that from now on, every judge will rule web crawling/scraping cases in the same way.

 **Reply**

Rick · 6 years ago

Brilliant bl~@dy aricle, and your a dev?

 **Reply**

Benoit Bernard · 6 years ago

Hey Rick - thanks for your kind comments.

And yes, I'm a dev :)

 **Reply**

Sergei · 6 years ago

Hi I still can't understand if crawling a list of publicly announced facts and arranging them in my way would be an infringement of the copyright of the original agenda website. If I don't copy the number of likes for each fact or the featured facts, or the way they are arranged, which I agree are the creation of the original website, then I am good, am I not?

 **Reply**

Benoit Bernard · 6 years ago

Hey Sergei - thanks for your question :)

First of all, IANAL (I'm not a lawyer). So I can't give you advice regarding your specific situation.

However, here are a few hints or clues:

1. Are you sure that you're just crawling "facts"?

2. If you look at the Linkedin cases given as examples in my post (Linkedin v. Doe Defendants), you'll see how things can turn out. Linkedin profile pages are essentially made of facts, but they're creatively arranged, so copying them in some way - like in memory, which is by definition necessary when crawling/scraping stuff - might be claimed as being a copyright violation.
3. The nature of your question shows precisely what the problem is. There's not a clear boundary about what will be tolerated, and what won't. Crawling/scraping stuff isn't illegal by itself, but what people **might** subsequently attempt to stop you from doing so is the **real problem**.

In case of doubt, please refer to the "General advice for your scraping or crawling projects" section of my post.

Good luck!

 Reply

Anonymous · 6 years ago

Hi, thanks for your article.

I have a similar problem to what some of the people in the comments described.

Basically I'm making a browser extension that will quickly find the first not-sponsored product on a site (using a scraper). This is used to bypass the first x pages full of sponsored products, to get to the 'real' listing of the products.

If many people used my extension, then the loss for the company could be very big. And I wonder how much the company or the law could hold me accountable for their loss of.

I get that it depends on them and how far they're willing to go. But I wonder if they could actually make a case that they lost a lot of money because of me and now I have to pay it back.

Any of your thoughts welcome.

Reply

Benoit Bernard · 6 years ago

Hey - thanks for your question.

Let me summarize what you want to do:

1. You want to create a browser extension that "skips" the sponsored products on a website.

2. You expect it to be popular.
3. You expect it to (financially) harm the company in question.

What you want to do is probably useful from a user's perspective, but it's also probably at odds with the company's objectives.

If I said that you should ask for their permission, would you do it? If your answer is "no", then you have your answer.

All in all, I suggest that you:

1. Read their ToS (Terms of Service). They probably have a section indicating what you're allowed to do with their website.
2. Consult a lawyer.

-1 | **Reply**

Souk · 6 years ago

I think this topic is mostly about indexing or scraping data from a website. What about indexing forwarded email data? What can be said about that?

| **Reply**

Benoit Bernard · 6 years ago

Hey Souk - interesting question, although it's not directly related to web crawling or scraping.

For your information, email data is copyrighted too. So you can't do whatever you want with it. See this **very interesting article** on this topic.

 **Reply**

Matthieu Harbich · 6 years ago

Hey, thanks for your article. Sorry if I missed the information (I didn't went through all the comments) but the examples you are using are facebook, Tweeter or LinkedIn. And for sure all those applications for which you need to login and thus accept conditions, you're not allowed at all to scrap the content, and use their API.

Is there not a big difference between a complete open website and a website on which you need to login and accept conditions?

Also, you mention that it's the "creative" part that can be protected, but that the facts that we want to store aren't. Am I right? When we crawl, we don't copy

the .css and we just care about specific markups, which are facts.

What do you think?

Thanks for the article.

M.

Reply

Benoit Bernard · 6 years ago

Hey Matthieu - thanks for your questions.

> *Is there not a big difference between a complete open website and a website on which you need to login and accept conditions?*

Well, at least concerning copyright, there's no difference. Websites and their content are generally copyrighted.

Now, as explained in my post, there are clickwrap and browsewrap contracts. The former means that you need to explicitly 'agree' to the contract's terms. The latter means that just using or accessing the resource makes you implicitly agree to the contract's terms. These kinds

of contract have been both enforced and rejected by different courts over the years. So it's up to you to decide.

> *When we crawl, we don't copy the .css and we just care about specific markups, which are facts.*

When crawling, you certainly copy the HTML data, don't you?

It turns out that CSS and HTML files aren't facts; they're both a creative arrangement of data which includes - among other things - facts.

For example, a price or a date is a fact - it's not copyrightable. But a page (including its markup) in which these facts are included is generally copyrighted, because it's creatively arranged.

-1 | **Reply**

monu · 6 years ago

I want to use webhose.io API for news app and come to know that it crawls through many websites and provide me content.

But I want to know is it will be illegal if I show news of other news sites in my app?

Reply

Benoit Bernard · 6 years ago

Hi!

Content indexed by webhose.io - and more specifically "news" - is certainly copyrighted, so you most likely can't republish it without obtaining the original author's written permission.

I suggest that you read their ToS to verify: https://webhose.io/tos/. In case of doubt, contact their customer support team, and possibly consult a lawyer, especially if you intend to make money with your app.

Good luck!

1 Reply

monu · 6 years ago

Thanks for the reply...It was helpful.....Do You know about YouTube API...I have some questions if you could answer?

Reply

monu · 6 years ago

YouTube provides API for play video in android app. But should I have to get permissions from channel owners before play their videos in my App?

```
The main functionality of my app is not YouTube video so,

(i) Can I monetize my whole app except YouTube Video Activity
(ii) Can I monetize my whole app including YouTube Video Activity

(iii) I can't monetize my app
```

Please provide me answers of two these question

Reply

Benoit Bernard · 6 years ago

Hi Monu,

Your questions are about APIs, not about web scraping or crawling. Those are quite different. I haven't sufficiently researched APIs to know exactly what legal aspects need to be considered. Also, I'm not a lawyer.

Now, based on common sense and on what I know about web scraping/crawling, here's what I'd personally do:

1. Read Youtube's ToS. Maybe they prevent you from including videos in an external app, or maybe they prevent you from making money that way.
2. Ask the authors of the videos for a **written permission** to use/republish/monetize them. Those videos are almost certainly copyrighted.
3. Consult a lawyer, especially if you intend to make serious money with your app.

Also, concerning your questions about monetization:

> *(i) Can I monetize my whole app except YouTube Video Activity*

Well, I guess that it really depends on **how** you intend to monetize your app!

 Reply

Juan · 6 years ago

The LinkedIn vs HiQ court case is not referenced yet. Any update in the

situation exposed in the post?

Reply

Benoit Bernard · 6 years ago

I haven't researched that court case specifically, but you can watch the hearing **here**.

Reply

Anthonus · 6 years ago

After 35 years of building businesses worldwide and working in many countries (being involved in all aspects and politics), my VERY HUMBLE opinion is that "the old boys' club" prevails! If you are not part of the 'hayward-poison-ivy' club you are doomed! In other words 'some are more equal than others'.

Big companies belong to the 'old boys' club :)

Legitmacy has NOTHING to do with it - I agree - it is about 'who you know and not what you know'... in other words 'The Internet' is an 'elitist' playing field...

Yep, big companies can do whatever it wants - it is a 'big boys'' game... do a background research in to all those startups and 'I bet' you will find the 'gravy train'... sadly, nothing is for the 'commoners'...

-1 | **Reply**

Benoit Bernard · 6 years ago

Hey Anthonus - thanks for your comments.

Early innovators (Google, Facebook, Linkedin, Amazon, etc.) were able to capitalize on the fact that they were all alone back in the day - and could consequently harvest as much web data and personal information as they wanted. At that time, nobody bothered. And for that reason, they were able to get an immense advantage over their (future) competitors.

Then, those early innovators realized that to keep their dominance, they needed to protect their data - hence the protection measures that they put in place (e.g. LinkedIn having anti-scraping systems). In addition to that, some laws emerged in the past couple of years to further protect the privacy of people (e.g. GDPR).

So right now, it's certainly harder than ever for new entrants. I mean, at least for those wanting to enter the same markets.

| **Reply**

Anthonus de Silva · 6 years ago

Hi Benoit,

Thank you for accepting my comment. I agree with what you are saying, but what I am 'insinuating' is that the playing field was 'preordained' (the stage was set and suitable 'characters' were 'selected'). For example, Hotbot was a better search engine, Hi-5 was there way before Facebook and we have all heard of the 'rumours' sorrunding the 'Facebook/Hi-5' https://www.quora.com/Did-Mark-Zuckerberg-steal-the-idea-for-Facebook-from-the-Winklevoss-twins-considering-they-did-get-65-million-from-a-settlement-last-year-How-legitimate-is-their-claim). Facebook (seemingly) blocks many accounts that are critical of 'western' governments. but allows for 'propaganda' accounts... (http://www.wicipolskie.org/?p=23940 and https://www.reddit.com/r/conspiracy/comments/2geokf/mark_zuckerberg_is_jacob_greenberg_is_grandson_of/). Whatever the truth is, Mr. Zuckerberg was/is definitely not a 'boy in a hoodie' who invented a 'web site' and struck it rich! No such thing exists in the world.

Then we have the Bill Gate's mother's IBM and governor of Washington 'rumours' (https://www.nytimes.com/1994/06/11/obituaries/mary-gates-64-helped-her-son-start-microsoft.html). You may have

already heard that 'Garry Kildall accused Bill Gates of stealing CP/M'. I know there is some credibility as I programmed in CP/M, and when the BBC asked Mr. Gates, sometime in the late 1980s, 'What does the erase command in MS-DOS do?', Mr. Gates replied 'MS-DOS does not have an erase command'. But if you load an early version of MS-DOS, you will see that MS-DOS, does indeed, have an 'erase' (or could have been 'delete') command (CP/M has an erase command). Then Digital Research (The company of Garry Kildlall) produced GEM (Graphic Environment Manager) - Looks a lot like Windows 2/3 and Apple's LISA...

To top it all. GDPR is not new. All my (businesses) IT work since 1989 have been compliant with the 'British Data Protection act'. In other words we are now (automatically) GDPR compliant. In this light, Canadian and US companies are still finding ways to 'circumvent' GDPR by things like 'accept our cookie policy' (GDPR simply means that 'you can't profit with user data without their express permission' - so the cookie policy should read 'Please note that we will be selling your browsing habits to advertisers and companies - you privacy is compromised') . Europe is way ahead in the world with protecting consumer rights.

The questions to ask these days is 'Does Mr. Trump own shares in Twitter?', or, 'Did Mr. Obama own shares in Blackberry or have some stake in the Canadian city Mississauga', etc.

You may not wish to publish my comment, and I won't hold it against you. I am just sharing my 'experiences' with you as could use my knowledge/experiences to good use. However, if you do publish my comment, it just might help make the world a better and fairer place...

Thank you.

Reply

Benoit Bernard · 6 years ago

I feel that we're deviating a little bit from the topic at hand - web scraping and crawling. But your comment is interesting and kind of related, so let's see.

> *What I am 'insinuating' is that the playing field was 'preordained' (the stage was set and suitable 'characters' were 'selected').*

Personally, I don't feel that the playing field was "preordained".

On the contrary, I think that the market decided what it wanted. And I think that Microsoft, Facebook, Google, Twitter and LinkedIn were just better than their competition.

By *better*, I mean in terms of idea, product, business model, execution, management, marketing/PR, connections, ethics (or lack thereof...), etc.

> *Canadian and US companies are still finding ways to 'circumvent' GDPR by things like 'accept our cookie policy'*

Indeed, I've seen many Canadian and US websites offer users to accept their cookie policy. But many of them don't provide an obvious and easy way to opt out.

Is that what you're referring to when you say that companies are trying to circumvent GDPR?

> *Europe is way ahead in the world with protecting consumer rights.*

I've heard (and read) the same thing.

> *The questions to ask these days is 'Does Mr. Trump own shares in Twitter?'*,

> *or, 'Did Mr. Obama own shares in Blackberry or have some stake in the Canadian city Mississauga', etc.*

I understand your point, but I really don't think that Trump owns shares of Twitter. Twitter was not even invited to Trump's Tech Summit.

If they were to ban Trump from their platform, he'd probably find and use another one. So instead of losing that traffic, it's probably just better - from a business perspective - to keep him.

Reply

Anonymous · 6 years ago

Compounding the problem is one can higher an army of international labor for cents on the hour to simply manually 'scrape' websites. Automatons mindlessly keying in web addresses and copying the pages provides the same capabilities of a robot. Taking the idea a step further...one could build a theoretical robot (I'm talking about a physical robot here) to 'manually' type the commands into a keyboard. Wonder how site's ToS would deal with that...let alone the courts #ROTFL

Reply

Benoit Bernard · 6 years ago

I agree - there's probably always a way to circumvent the law or any security measures put in place by a company (e.g. LinkedIn).

Having said that, how scalable and realistic is the approach that you described? Moreover, in my mind, hiring an army of international labor or using physical robots clearly falls into the "automated" kind of use that is generally prohibited by ToS. Also, the content that you'd "manually" scrape is most likely copyrighted.

 **Reply**

Anthonus · 6 years ago

Compiling (taking) the data is NOT the problem - how you use it is! Copyright law and Plagiarism are two different things. but only a judge can decide. I have seen judges decide on if there was a copyright violation or not based on the 'originality component' of the product. A general rule of thumb in what I witnessed was that at least 30% of the work must be original (this is why, for example, Linked-In, blocks scraping....).

Lawyers will try to arbitrate (bully?) into submission. But if you believe

your work is original, irrespective of scraping/etc, then you can state your case - and - if it goes to court the judge will 'decide' - but, sadly, there may be other (political/social/re-election) implications, so the judgement may not be what you expect...

I have developed IT/Tech since 1980 (professionally since 1986). I value original work, I use 'Open Source' but work extensively on the product(s) to make it 'original' (only use Open Source as a RAD tool - the last time we downloaded an Open Source product was 2010 - even the CentOS versions we run now are customized/debugged and worked on by our people), etc.

To be frank, I really believe (experienced) that if someone believes that they can make a 'fast buck' out of simply scraping, they are mistaken (at least it is not nice). Try to do something original and you will get 'lucky'. I am Roman Catholic... a Buddhist saying goes 'Pure Intentions will pave the way for good fortune..'. In that light, I have done business since 1989, my intentions have always been good and I have survived 'the world' of lawyers and judges...

Bottom line: If you want to do any kind of business, it would be advisable to study up on 'your/applicable jurisdiction's' trade/business

law. Back in the 1980s I had to do it the 'hard way', today you can just do an internet search such as 'California copyright law', etc.

On a side note: I don't do social media. I am too busy for that running a small business in 3+ countries (I have CEOs for all of them)... in that light, I really don't believe Mr. Trump does twitter by himself... the misspelling of 'precedent' was how it was 'legitimized' to the public.... Mr. Trump and his family are very successful, would you really believe that Ms. Ivanka Trump personally does 'twitter' or even the 'Kardashians' don't hire people to do social media?

I am 52 years now, never met anyone who made a 'fast buck'... making 'a buck' is a lot of work... people I met over 30 years ago who tried to make a fast million are still trying or worse...

As someone said 'success is 95% perspiration and 5% inspiration'.

I trust this helps. I don't believe it is off topic.

^ | v **Reply**

Benoit Bernard · 5 years ago

Hey Anthonus - thanks for your comments.

> *Compiling (taking) the data is NOT the problem - how you use it is!*

You're wrong about that. Taking the data and using it can be both problematic, as shown by the various court cases included in my post.

For example, LinkedIn didn't even know what those 100 anonymous people intended to do with the data that they scraped. And yet, LinkedIn sued them.

> *Lawyers will try to arbitrate (bully?) into submission.*

Yes, you're absolutely right.

> *But if you believe your work is original, irrespective of scraping/etc, then you can state your case - and - if it goes to court the judge will 'decide' - but, sadly, there may be other (political/social/re-election) implications, so the judgement may not be what you expect...*

By the time you go to court, it might be too late. How much money will you have spent in legal fees?

> *I really believe (experienced) that if someone believes that they can make a*

> *'fast buck' out of simply scraping, they are mistaken [...]*

I agree.

> *If you want to do any kind of business, it would be advisable to study up on 'your/applicable jurisdiction's' trade/business law.*

Excellent advice.

Reply

LR · 6 years ago

Fabulous post! If only everything on the 'net was as well thought out and presented.

Although it is a bit rich that Google (Of all people!) can prosecute crawling & scraping in some ways the restrictions don't seem totally unreasonable - you wouldn't expect to be able to extract content from library books without infringing copyright. Whether or not it is 'right' that Google and others can enforce a constraint on the openness of the 'net that they themselves have exploited so profitably, is quite another matter.

My guess is that legal action will only occur when the plaintiff feels it is

worthwhile i.e a likelihood of success worth the effort & cost of the legal action. With so many crawlers and scrapers in action (Those who know the game, and the many more who don't) the odds are good, perhaps overwhelmingly so, that an individual case will escape the notice of google/fb & the like.

Reply

Benoit Bernard · 6 years ago

Hey LR - thanks for your kind words :)

> *Whether or not it is 'right' that Google and others can enforce a constraint on the openness of the 'net that they themselves have exploited so profitably, is quite another matter.*

Yep, this is unfortunately the reality that we live in.

> *My guess is that legal action will only occur when the plaintiff feels it is worthwhile [...]*

Precisely. If you pose a threat to them, then they'll do something about it.

> *the odds are good, perhaps overwhelmingly so, that an individual case will*

> *escape the notice of google/fb & the like.*

While this might be true in most cases, there are many cases where things didn't turn out as expected. Think about Pete Warden and Facebook, for example.

This is probably comparable to Napster and the likes back in the day. Most people downloaded music for free on those services, and only a minority got caught/targeted by the RIAA. But they still got caught, and they paid the price for everyone else.

 Reply

Joshua Blue · 5 years ago

So basically, if I am on rotating residential proxies. Am I breaking the law? Even though it seems kinda impossible to track via back-connect proxies, is that correct?

 Reply

Ben Bernard · 5 years ago

Hey Joshua,

The very act of scraping/crawling isn't illegal by itself.

However, I suggest that you carefully read the terms of service of the website that you intend to scrape/crawl. Pretty often, there's a clause about "automatic data collection" that prevents you from doing so.

So no matter what method you use to scrape or crawl - be it with a scraper running on your laptop or with a distributed web scraper running behind residential proxies - you may still be legally bound to the website's ToS.

Sure, they'll most likely fail to track you down. But hey, the final decision belongs to you :)

Reply

Jamie Leighton · 5 years ago

When Facebook scrubs user information are they required to send a legal copy of the information to the Federal Government as record?

Reply

Ben Bernard · 5 years ago

Hey Jamie - to be honest, this isn't my area of expertise. But based on **this page**, I suppose that any law enforcement official - at least one who is authorized to do so - is able to request user data.

Reply

Riccardo Esclapon · 5 years ago

You mention the case LinkedIn v. Doe Defendants as proof that you shouldn't do web scraping, but the outcome was in favor of the Defendants: http://www.baercrossey.com/1982/federal-district-court-ruling-against-linkedin-has-major-implications-for-web-scrapers

Reply

Ben Bernard · 5 years ago

If you pay close attention, you'll find out that these are completely different cases. You're actually referring to LinkedIn v. hiQ, not LinkedIn v. Doe Defendants.

For more information about hiQ, see one of the comments above - we've already discussed it.

Reply

A · 5 years ago

Does anyone here have pointers to info on legality of the practice of crawling the web searching whether their own data is being published

elsewhere without authorization? Let's say the script would just crawl and create two lists - one with links to pages not including the data and one with links to pages that seem to include it (Which could then be visited by a human). No other data except the links would be stored.

Reply

Ben Bernard · 5 years ago

Hey - thanks for your question.

So you want to scrape let's say Google's search results to see whether someone else has republished your content on their own website?

Well, if you look at Google's ToS, you'll see that they prevent "automatic data collection" of their search results.

Reply

Anonymous · 5 years ago

Hey Ben, thank you for ur info. It is pretty helpful to clear out some of the grey area for the legal side. I still have a few questions regarding to the article: How do you determine if the scraping/crawling require permission? Does it mean that if the website doesn't include Terms of Use then we should be good to go? How about the website that said the content is under DMCA? Does

scraping/crawling video from video website (Like YouTube) have different problem?

Reply

Ben Bernard · 5 years ago

> *How do you determine if the scraping/crawling require permission? Does it mean that if the website doesn't include Terms of Use then we should be good to go?*

It really depends on the website itself.

A good rule of thumb is to:

1. Assume that a website always requires permission to be crawled or scraped.
2. Assume that the content is copyrighted.

Look at their Terms of Service (ToS), or check if there's any directive in their robots.txt about scraping/crawling.

If you find nothing, ask the website's author for a **written** permission to

crawl/scrape it and/or reuse their content.

> *How about the website that said the content is under DMCA?*

Well, you mean a website with a DMCA badge? In any case, this means that the website's content is copyrighted, and that if you republish it somewhere, you might get a notice to take it down. If you don't take it down, then well, things might go further.

> *Does scraping/crawling video from video website (Like YouTube) have different problem?*

Again, the content is almost certainly copyrighted. And YouTube's ToS (see here) explicitly prohibits you from using "unauthorized means" to access their content.

This means that you can't write a script that automatically downloads videos from YouTube.

Reply

Jordan Hansen · 5 years ago

REALLY great article and discussion here.

Do you feel that the result of the linked in case would change your opinion here at all? https://arstechnica.com/tech-policy/2017/08/court-rejects-linkedin-claim-that-unauthorized-scraping-is-hacking/

Reply

Ben Bernard · 5 years ago

Hey Jordan - interesting question.

As I understand it, the case is still running its course. LinkedIn can still go in appeal, so nothing is final yet.

And I don't think that anything in this specific case prevents LinkedIn from threatening other companies that scrape their user profiles.

2 Reply

Anonymous · 5 years ago

ToS can't be enforceable if it prohibits activities not prohibited by the law. ToS can prohibit you to breath while you are using the website. I am not saying that you don't have points but the weighted arguments should be on the law/sue cases side, not the ToS.

Reply

Ben Bernard · 5 years ago

This is incorrect.

Please read bullet #8 above (about the enforceability of ToS).

Reply

katsarov · 5 years ago

Thanks for your post! Currently I have made a price scraper for eshops providing users a easy way to follow favorite products and their prices. And I want to rise a question which I didn't find in comments: What about microdata and json-ld? They are made for robots. The store on purpose provide their info to robots.

Reply

Ben Bernard · 5 years ago

Hey - interesting question.

Microdata and JSON-LD are made for search engines and SEO. Now, if you're wondering whether your price scraper - which is essentially an automated script - can scrape that information, then I'd suggest that you read the website's ToS and see whether they allow any form of

scraping. If it's unclear, then contact the website's owner.

Of course, I'm making these suggestions based on the assumption that you'll want to make money out of that scraped data (and/or that you'll heavily poke the website in question), which might not be the case at all. In any case, use your judgment :)

Reply

katsarov · 5 years ago

The first reason I made it was because some websites rises the price before a promotion and they cheat. So - probably will not be very happy with that :) Technically the project is some kind of search engine (for products). The rate is very low - number of categories * number of pages per 24 hours. It shouldn't be big deal for them. But yes - the line is thin. Will see :)

Reply

Ben Bernard · 5 years ago

If I were you, I'd be careful. If your search engine gets quite a bit of traction and they discover it, they might not like what you do (especially if it's against their ToS). I'm just saying.

1 Reply

Jon D · 5 years ago

Thanks for this article. I found it very helpful to understand the legal aspects of this activity. We have started to read supplier websites terms and conditions with a little more enthusiasm!

One query. We do send out to our customers (consumers) personalised emails which can refer to product information on stockist websites (usually just a link to a specific product page). We have noticed in terms and conditions phrases like "You may not create a link to this website from another website or document without our prior written consent". Any views on reasonable ways to address this? If we cant reproduce the material (I understand why) and we cant provide a link to their website I am at a loss what to do. I suppose we could say search google for "xxxx" but it seems pretty customer unfriendly.

-1 | **Reply**

Ben Bernard · 5 years ago

Hey Jon - my pleasure.

> *Any views on reasonable ways to address this?*

Their terms are pretty clear - they don't want you to link to any one of

their product pages. They probably have good reasons for it.

But I'm wondering, what are you trying to do exactly? Are you scraping product pages to find good deals for your email subscribers?

If so, then they might not like it... unless they see what you're doing as something benefiting both parties.

Now, I don't know how enforceable a "no linking" clause might be, but since your business seems to depend on it, I'd suggest that you:

1. Ask for their written permission.
2. Seek advice from a qualified lawyer.

Reply

Jon D · 5 years ago

Hi Ben, thanks for that. On reflection I agree that written permission is the only way to go here.

We provide guidance to consumers on what an overall product/service package should cost (eg gas central heating installation) so use scraping

to check that our guide prices on specific product elements (such as boilers) are accurate and we are actually presenting current rather than obsolete products. We try and publish "fair" prices rather than promote the cheapest but thought it would be helpful to suggest a few online outlets where this product could be bought at around the price we suggest, when a consumer requests we email them the details (this is an automated process). It should be easy enough to get agreement from some of the online retailers and just use the others as a sanity check. Very helpful to clarify our thinking. Kind regards Jon

Reply

Noel Rolls · 5 years ago

Can I ask your opinion on whether web scraping a number of commercial property websites then aggregating and geocoding the data to get an index for various localities would be questionable? The idea is that people could compare a property they are interested in to the index (in terms of cost per unit of space). Thanks

Reply

Ben Bernard · 5 years ago

Hey Noel - thanks for your question.

Again, this boils down to whether or not you respect their ToS (Terms of Service), and whether or not you're copying copyrighted content.

If I were you:

1. I'd research this court case: **Craigslist Inc. v. 3Taps Inc.**
2. I'd consult a qualified lawyer, especially if you plan to make money with your idea and/or compete with the commercial property websites.

Good luck!

Reply

Jeppe · 5 years ago

If it is publicly accessible - I do not really see why it would be illegal. You can bypass robot.txt with any residential proxy network and do whatever you want. It is like setting up an information poster, but not letting everyone look at it.

Reply

Ben Bernard · 5 years ago

Hey Jeppe - let me address your points.

> *If it is publicly accessible - I do not really see why it would be illegal.*

What about copyright laws?

By the way, I've never said anywhere that scraping by itself is illegal.

The whole point of my post is that scraping/crawling is not illegal, but that you can still get in trouble for doing it (if you're not careful enough).

You seem to be assuming that because something is not illegal, you can't end up in litigation over it. In fact, you can end up in litigation over pretty much anything - especially if a person or a business thinks that you've caused them damages.

Finally, if you ever end up in litigation over your scraping activities (which is unlikely, but possible), having used a proxy network won't be very effective in showing your "good faith" :)

Reply

Anonymous · 5 years ago

Nice 3 year discussion going on here still alive and kickin! Great long post about all this by the way.

On a smaller scale, what is your take on crawling local business websites for information gathering with the intention to send them an email to conduct B2B services?

What about a different scenario where you outsource a VA to manually visit each website and extract contact details 1 by 1?

There are people who scrape and send millions of emails and there are people who scrape and just 100 emails, yet it sounds like the law can land upon anyone no matter how big or small or does it just land on the big guys?

Reply

Ben Bernard · 5 years ago

Hey - interesting questions.

> *what is your take on crawling local business websites for information gathering with the intention to send them an email to conduct B2B services?*

Are you requesting web pages faster than a human would? Are you respecting their robots.txt? Are you respecting their ToS? Are you infringing on their copyright in some way?

A first step would be to answer these questions. Now, I know that many businesses and startups are already doing this. Personally, I don't see it as either right or wrong, but you should make sure that you understand what you're doing beforehand.

> *What about a different scenario where you outsource a VA to manually visit each website and extract contact details 1 by 1?*

At least, a website owner couldn't claim that you used an "automated tool" to extract information from their website :)

In their eyes, this would most likely appear as a regular guy browsing their website... unless you ultimately publish the data that you extracted.

> *[...] it sounds like the law can land upon anyone no matter how big or small or does it just land on the big guys?*

Unfortunately, there's no absolute answer to your question. It appears to work on a case-by-case basis.

Reply

Steven · 5 years ago

I have been using RTC tool by Oxylabs to mainly scrape e-commerce websites, thought there wasn't anything wrong with that as all the data you would scrape is already publicly available so as long as you follow the robots.txt rules for a given website you should be in the clear right?

Reply

Ben Bernard · 5 years ago

Hey Steven - thanks for passing by :)

No matter what tool you use, the potential concerns stay the same.

Please refer to this section to get an answer to your questions: **The typical counterarguments brought by people**.

Reply

---

jocke · 5 years ago

Hi Benoit Bernard. This is a very interesting article. A question or two:

If you would scrape let´s say a couple of thousands induvidual FB, IG or Twitter accounts and get snippets and present them top to bottom on your own website... If that was done...would it make any difference if you had a written permission of the individual owner of the account at any of those places? Or is the conflict still with the service? I am thinking that the individual person is providing text/photo to the....Facebook account for instance. I then use a scraping service...If I have permission from the original poster to use that post to get the info/text/photo to my website...am I "tresspassing" still? In legal terms...?

Reply

Ben Bernard · 5 years ago

Hey Jocke - I'm not a lawyer, so I can't give you advice regarding your specific situation.

But I would guess that you still need to get a written permission from the profile's owner **and** from the website.

The number of profiles that you scrape doesn't matter at all.

Reply

Cst · 5 years ago

I was wondering... If I subscribe to a service and log into a Web page that delivers live data... If I capture and filter out data, that are requested from the Web page (running in my Browser), and use that data to derive my own data, would that be considered scraping? I am not talking mirroring the data to other people (which would violate the terms of my subscription), but using the derived data. A simple example: I get the information "RED" and "BLUE" coming in from the Web server and I mix the two colors and come to the conclusion that the resulting color is "VIOLET". I guess another of the "Grey zones"??

Reply

Ben Bernard · 5 years ago

> *If I capture and filter out data, that are requested from the Web page (running in my Browser), [...]*

How do you plan to "capture and filter out" the data? Using an automated tool? If so, then check the Terms of Service of the website in question - it may prohibit any form of "automatic data collection".

> *and use that data to derive my own data [...]*

I'd suggest that you verify under which license that data is made available. There may be restrictions put in place.

Reply

Anonymous · 5 years ago

Capturing can be done in two ways:

1. Doing the same requests that the Web page is doing, in the same or lower frequency.
2. Capturing the packages on MY network

The data that I'm talking about is something that is picked out of the air (aka 'public' data), and I use (and pay for) the 'service' to forward the data from locations where I cannot collect the data myself. So, not proprietary data.

I have no use of the data in the form it is presented on the Web page, and as such I turn it into MY 'proprietary' data.

Some of the service providers are adding disclaimers along this line:

"Some data is provided by 3rd parties and we cannot guaranty the quality and validity of the data and accept no liabilities".

So am I allowed to analyze the data and figure out if it is 'junk' data?

I have a lot of "automatic data collection" software on my computers, like spam filters and virus protection. They all do "automatic data collection" in order to protect my computers and filter out annoying data?

Like I stated, probably one of the 'Grey areas'???

Reply

Ben Bernard · 5 years ago

> *So am I allowed to analyze the data and figure out if it is 'junk' data?*

I seriously don't know!

As I said, check the license and the ToS. And ask for a written permission to use their data if you're still not sure.

What you see as "public" may not be as public as you think it is - at least not from the eyes of the company whose data is being scraped.

> *I have a lot of "automatic data collection" software on my computers, like spam filters and virus protection.*

Just to be clear - what I really mean is "automated data collection", not "automatic data collection". Sorry if it confused you.

Now, I don't understand your point. Are you assuming that because there's software that performs some form of automated data collection on your PC that it automatically grants you the right to perform any form of automated data collection on websites that you don't own?

If so, then it's a mistaken assumption.

1 | **Reply**

Anonymous · 5 years ago

Maybe I should clarify: I have a subscription that allows me to use the data commercially. I get access to the data over a Web page, but presented in a format that is useless to me. The Web page is requesting data from the server every 5 seconds. Instead of using the Web Page I

request the data from the server with an identical query (but at a lower rate) and then use/present the data in my own format. So I'm not overloading their server (on the contrary, because my subscription allows me to have 3 Web pages open at the same time), and the data is being used internally and is not being resold. (But never mind. We have found a better and more reliable way of getting the data that we need and might even consider to turn up as a competitor with our OWN data in the future)

Reply

David · 5 years ago

Hi Bernard,

Thanks for this most interesting article, even three years after its date of publishing. I am new to email marketing and scraping and all those tactics/actions/techniques. I have gone through all the comments to find an answer to my question but haven't seen it. A comment you have given to an anonymous person four months earlier came close, but I hope that you don't mind me asking?

My question is: if I (done by myself, not a VA or agency from abroad) manually "scrape" a website for the email address, would it be possible to get in trouble?

I will use the email address (and only the email address) to send an email that asks for permission sending the reader a video about something regarding their business. The reader has to opt-in.

I am still considering if I should use double opt-in or if a single opt-in will do. If the reader opens the email but chooses not to opt-in, they will not hear from me again. If they don't open, I might send another email next month.

I do not use any of their copy of their site.

I state clearly in my mail that it is my attention to get into a mutually beneficial B2B relationship.

Each mail has a clearly visible "unsubscribe" button.

I will send a max of 20 mails (so 20 different receivers) per day.

Their TOS might forbid using their email address. I don't know. I prefer not to read/study all their TOS's to find out.

What is your take, Bernard? Do I put my business, in danger by just scraping the email address? And if I do, how much of a difference would it make that I am

operating from an LLC?

1 | **Reply**

Ben Bernard · 5 years ago

Hey David - these are difficult questions.

> *if I [...] manually "scrape" a website for the email address, would it be possible to get in trouble?*

Theoretically, yes.

But concretely and practically, unless the website's owner figures out what you're doing and they're motivated to stop you, the answer is probably no.

But this depends on so many factors and I'm missing a lot of context and details here.

You shouldn't take the above as any kind of legal advice - I'm not a lawyer.

> *Each mail has a clearly visible "unsubscribe" button. I will send a max of 20*

> *mails (so 20 different receivers) per day.*

There's nothing wrong with "cold emailing" prospects.

But what you're describing here looks a lot like spam. Why would people need to unsubscribe from something that they never subscribed to in the first place? You should be careful here - there are laws in some countries regarding spam, like in Canada.

> *Their TOS might forbid using their email address. I don't know. I prefer not to read/study all their TOS's to find out.*

Why?

If I were you, I wouldn't take any shortcuts.

> *Do I put my business, in danger by just scraping the email address?*

See my answer to your first question.

> *And if I do, how much of a difference would it make that I am operating from an LLC?*

This goes way beyond my area of expertise, unfortunately. You should seek advice from a qualified lawyer.

1 | **Reply**

Anonymous · 4 years ago

Thanks for the efforts you put in writing this article and answering to the comments. I rarely see bloggers answer to comments 3 years after the original post was written, I commend you for that.

I had a rather small personal project that involved scraping several websites of large international companies and maybe use their hidden but unprotected APIs. After starting playing around, I started wondering if doing this was legal and found your article. Needless to say, after reading it I changed my mind, it's not worth getting into troubles with that kind of people.

Anyways, my project sank to bottom of the ocean of abandoned projects but I learned very interesting stuff thanks to you.

| **Reply**

Anonymous · 4 years ago

We are getting in an era of scraping (need for data). HiQ vs Linkedin victory gives a lot of hope :) https://www.eff.org/deeplinks/2019/09/victory-

ruling-hiq-v-linkedin-protects-scraping-public-data

Reply

Nico · 4 years ago

Ben, thanks for all the research and efforts you've put into writing this blog and answering all the comments! I'd like to hear your thoughts on this one:

What if you had website with an input field that asks the user to submit their website's url which you then scrape to assess it in terms of usability. A lot of "SEO tools" do that.

Q1: Technically YOU are scraping the website. So are you or the user who submits the URL violating the TOS?

Q2: Would you log the user's IP and inform him upfront about possible legal actions in case he's not submitting his own website?

Q3: What's an easy way (possible non-technical) to validate a user's ownership of a website prior to scraping it?

Thanks, Nico

Reply

South Jersey says Don't Be Stupid · 4 years ago

This seems like a FUD article posted by a pro-corporate or pro-solicitor/lawyer lobbyist. Scrape ethically - limit your effect on the target site (aka "don't be an asshole") - but do NOT do something stupid that will identify you like a custom user agent that directs a potentially hostile party to your front door. Any lawyer with a brain will tell you the tell you the same.

I am not a lawyer either, but some of your advice is the kind that will make a lawyer have to get involved in one's day to day.

Its also now out of date, check out LinkedIn vs. HiQ. US court ruled collection of publicly available information is legal.

Reply

Ben Bernard · 4 years ago

> *This seems like a FUD article posted by a pro-corporate or pro-solicitor/lawyer lobbyist.*

Well, I'm neither of these things. I'm just a programmer who realized that there's much more to "just scraping a website".

> *Scrape ethically - limit your effect on the target site (aka "don't be an asshole") - but do NOT do something stupid that will identify you like a custom user agent that directs a potentially hostile party to your front door. Any lawyer with a brain will tell you the tell you the same.*

So everything's fine... until you get into trouble?

> *I am not a lawyer either, but some of your advice is the kind that will make a lawyer have to get involved in one's day to day.*

Absolutely. What's wrong with this?

> *Its also now out of date, check out LinkedIn vs. HiQ. US court ruled collection of publicly available information is legal.*

Are you sure? Please read the **EFF's article** carefully.

> *While this decision represents an important step to putting limits on using the CFAA to intimidate researchers with the legalese of cease and desist letters, the Ninth Circuit sadly left the door open to other claims, such as trespass to chattels or even copyright infringement, that might allow actors like LinkedIn to limit competition with its products. And even with this*

> *ruling, the CFAA is subject to multiple conflicting interpretations across the federal circuits, making it likely that the Supreme Court will eventually be forced to resolve the meaning of key terms like "without authorization."*

Reply

Tim · 4 years ago

Hello Ben, great write-up. Unlike most internet related writings, this article is on point and on-track with today's environment. Wondering if you have an opinion on how this relates the the recent revelations concerning the image scrapping efforts of Clearview? Perhaps you can re-post this article with an update, focusing on Clearviews' use and collection of images and development of their "for-profit" Facial Recognition technologies. Interested to hear your take on this. Cheers

Reply

Ben Bernard · 4 years ago

Hey Tim,

It looks like Clearview scraped a massive amount of Facebook, LinkedIn, Twitter, YouTube and Venmo profiles. This is clearly against their TOS.

In response, they apparently all sent cease-and-desist letters (*) to Clearview. But I don't know what the current state of affairs is.

Additionally, a **class-action lawsuit** was filed against Clearview in Illinois due to their non-respect of biometric state laws.

My opinion is that Clearview either knew exactly what they were getting into from a legal perspective, or they didn't :)

*: **Facebook**, **LinkedIn**, **Twitter** and **YouTube**

Reply

Anonymous · 4 years ago

Do you intend to update this article to reflect the ruling in LinkedIn's case (https://www.forbes.com/sites/emmawoollacott/2019/09/10/linkedin-data-scraping-ruled-legal/#47062e481b54) which clearly renders much of the analysis in this article null and void?

Reply

Ben Bernard · 3 years ago

I think that you're declaring victory a little too soon.

Have you read **EFF's article**?

> *While this decision represents an important step to putting limits on using the CFAA to intimidate researchers with the legalese of cease and desist letters, the Ninth Circuit sadly left the door open to other claims, such as trespass to chattels or even copyright infringement, that might allow actors like LinkedIn to limit competition with its products. And even with this ruling, the CFAA is subject to multiple conflicting interpretations across the federal circuits, making it likely that the Supreme Court will eventually be forced to resolve the meaning of key terms like "without authorization."*

Reply

Mark Anthony Carrillo Sr · 4 years ago

It's legal but should not be allowed to remove information that was already public....

Reply

Ben Bernard · 3 years ago

I disagree with you concerning the second part of your statement. In my opinion, you should be freely able to either remove or make private any of your personal information that is currently public.

Reply

Vance · 4 years ago

This is an exceptionally informative and well-written post, thank you so much for sharing your research on this and taking the time to include all those specific case references! Answered pretty much every question I had on the matter. Just one thing I'm curious about: do you have much experience with actually contacting a site, explaining what you want to scrape their data for, and asking permission? Do you have a sense of how well or poorly folks tend to take that? (assuming of course that your purposes aren't blatantly against their interests)

Reply

Ben Bernard · 3 years ago

Hey Vance, thanks for your kind words!

I don't have much experience with asking for a written permission to crawl or scrape a specific website, but I do have experience with setting up a page that announces what my crawler is doing (and why) and linking to it in my user agent string. Creating such a page will prevent many misunderstandings and potential issues.

For example, I set up such a page for the big crawling/scraping project that I'm referring to at the beginning of my post. According to my blog's analytics, it worked wonderfully - following each crawl, many website operators visited the page, but I didn't receive any email, any complaint or cease-and-desist letter. The reason is probably because I was 100% transparent and people knew exactly what to expect from me.

For the actual content of the page that I linked to in my user agent string, here's what I included:

1. An email address to reach me directly.
2. The detailed purpose of my crawler.
3. Which politeness measures I put in place.
4. How to prevent their website from being crawled (including the IP address of the machines I was crawling with).
5. I also stated clearly that upon request, I would destroy any data crawled from their website.

Reply

Workata · 4 years ago

Great article.

 |  **Reply**

Anonymous · 4 years ago

Thank you for the great post. Wish I had seen it a few months ago ;-) Anyway, is it ok for me to publish the "code" for my scrapers publicly? I don't mean to publish the scraped data, just the code... so that anyone wishing to scrape the target websites, can do so at their own risk.

I want to publish the repo publicly on GitHub under an MIT license.

-1  |  **Reply**

Ben Bernard · 3 years ago

You're welcome!

Unfortunately, I'm not a lawyer so I can't advise you on that sort of thing.

But the following may be helpful to know:

1. For my own project, I intended to crawl/scrape a large number of

high-profile websites. Then, I discovered that I could get in trouble for doing so, so I set up a disclaimer page (that I linked to in my crawler's user agent string) mentioning that I wouldn't fully publish my source code. The idea is that I didn't want to encourage other people to do the same thing. This was a really personal decision, and I can't know for sure if it had any impact in the end. But hey, I've never received any complaint or cease-and-desist letter.

2. In the **Linkedin Corporation v. Michael George** case, the guy was permanently banned from LinkedIn for creating a scraper advertised as being able to scrape LinkedIn. Something similar may or may not happen to you - I'm just saying.

1 | **Reply**

Anonymous · 3 years ago

LinkedIn lost their court case

| **Reply**

Ben Bernard · 3 years ago

If you're referring to the LinkedIn v. hiQ case, then I've already replied to a similar comment above, which you can read to know my view on this. Basically, you're right, but I think that you **shouldn't**

celebrate too early:

> *While this decision represents an important step to putting limits on using the CFAA to intimidate researchers with the legalese of cease and desist letters, the Ninth Circuit sadly left the door open to other claims, such as trespass to chattels or even copyright infringement, that might allow actors like LinkedIn to limit competition with its products.*

Reply

Anonymous · 2 years ago

Brilliant post!

Reply

Benoit Bernard

Read more posts by this author.

Canada *https://twitter.com/mbenbernard*

Share this post

Subscribe to Benoit Bernard

READ THIS NEXT

The Tale of Creating a Distributed Web Crawler

Around 6 million records with about 15 fields each. This was the dataset that I wanted to analyze for...

YOU MIGHT ENJOY

The Case of the Mysterious Python Crash

It was almost 11PM. My distributed web crawler had been running for a few hours when I discovered a...