Web-scraping, Web-crawling, and APIs Guidebook

Overview

- Data copyright in the US
- Online Data
    - Web-scraping
    - Web crawling
    - APIs
- What Governs Online Content
    - Computer Fraud and Abuse Act
    - Terms of Service
    - Copyright
    - Trespass to Chattel
- Relevant Examples
    - Feist v Rural Telephone
    - Linkedin vs. HiQ
    - US vs Aaron Shwartz
    - Politiwoops
    - Recipeasly
    - Ebay v. Bidder's Edge
- Appendix Checklist

Introduction: Engaging With Online Content

When you think of online content, you might first think of the front-facing content you can see: text, images, and videos. There are a handful of considerations you must think about when using front-facing content, including who owns the content, the copyrightability of the content, and whether or not your use of the content would be a fair use. But sometimes, you want to use back-facing content, such as databases, metadata, or code, and using these types of content requires additional considerations.

The purpose of this guide is to go over some common engagements with online content that go beyond the use of front-facing matter on websites: the use of online databases or APIs and using web-scraping or web-crawling to obtain and use online content. The first section of this guide provides a brief explanation of each type of engagement, followed by a section describing the rules and regulations that affect these engagements. Finally, this guide closes with a checklist to start the process of working through your own intended use.

Types of Engagement

Web-Scraping and Web Crawling

Web scraping is using a program to gather content from a website. This is most commonly done on the HTML (which governs the structure of a website). Scraping the HTML of a website allows a user to download and organize the content of a website including links and data that it uses. Web crawling works similarly to web scraping; however, web scraping only scrapes one web page at a time, while web crawling automatically scrapes a web page and all pages that are linked to that web page.[1]

[1] Martin Perez, “Web Scraping vs Web Crawling: What’s the Difference? | ParseHub,” ParseHub Blog, February 3, 2020, https://www.parsehub.com/blog/web-scraping-vs-web-crawling/.

APIs and Databases

API stands for Application Programming Interface.[2] It has many different uses, but this guide will only discuss APIs that are meant to allow users to easily request and store a website's data through programming. Databases can, but do not always, take the form of an API.

Laws Governing Online Content

Although it is legal to use web-crawling, web-scraping, and APIs to gather data, there are limitations. The limitations typically fall within these four categories:

1. Computer Fraud and Abuse Act
2. Terms of Service
3. Copyright
4. Trespass to Chattel

This guide will go through all of these categories more in depth. But first, there are a few common practices that can help ensure your use is authorized:

1. **Does the website's Terms of Service forbid any of your uses?** If the terms say that reproduction or copying is not permitted, you cannot legally display any information you gather from the site. If the Terms of Service prohibits web-crawling or web-scraping, you should not use web-crawling or web-scraping on the site.
2. **Is the information private?** Determine whether the information you're using could be considered "private." Did you have to log-in to see the information? Is it personal in nature? If so, there may be additional restrictions on its use.

[2] Petr Gazarov, "What Is an API? In English, Please.," freeCodeCamp.org, December 19, 2019, https://www.freecodecamp.org/news/what-is-an-api-in-english-please-b880a3214a82/.

3. **Is there a "robots.txt" file in the code that stops web-crawling or web-scraping?** Laws surrounding web-crawling and web-scraping prohibit you from bypassing this code.[3]

If you answer "yes" to any of these questions, consider finding another source for the data. Answering "no" to all of these questions does not guarantee that your use is legal, though it does make it more likely that you are able to use the content. Below is more information and detailed descriptions of the laws that govern online content and how you can make sure your uses are authorized.

The Computer Fraud and Abuse Act (CFAA)

The CFAA was first enacted in 1986 and was last amended in 2008. The act "prohibits intentionally accessing a computer without authorization or in excess of authorization." Critics, however, note that it leaves the term "authorization" undefined.[4] While the original act was limited in scope, the scope broadened through amendments throughout the years.

Perhaps one of the most notorious cases involving the Computer Fraud and Abuse Act, *U.S. v Aaron Swartz* reads like a cautionary tale against strict web-crawling and hacking laws. In 2010, Aaron Swartz created a program that automatically downloaded articles from JSTOR, an online library of academic literature, which went against the policies of the site. In creating this program, he used code to circumvent the restrictions to downloads on the site. In 2011, a federal investigation was levied and, despite the fact that JSTOR expressed a wish to not press criminal charges, Swartz faced 13 federal charges and up to 50 years in prison. Due to the stress of the proceedings and threat of

[3] Martin Perez, "Is Web Scraping Legal? Explanation and Examples | ParseHub," ParseHub Blog, July 14, 2021, https://www.parsehub.com/blog/web-scraping-legal/.
[4] "NACDL - Computer Fraud and Abuse Act (CFAA)," NACDL - National Association of Criminal Defense Lawyers, accessed July 29, 2021, https://www.nacdl.org/Landing/ComputerFraudandAbuseAct.

jail time, Swartz committed suicide in 2013.The government dropped the charges against Swartz after his death[5]

In the wake of Swartz's death, many internet freedom advocacy groups spoke out against the CFAA and its "draconian" enforcement. The EFF pointed to the lack of clarity within the CFAA and "heavy handed penalties" in enforcement of the CFAA in explaining what went wrong in the case.[6]

In more recent cases, such as *Van Buren v. United States,* the government has eased up on the enforcement of the CFAA and vague phrases such as "lack of authorization." In this *Van Buren*, a police officer, Nathan Van Buren, used his credentials to run a license plate in exchange for money. This action was a violation of policies in law enforcement, and he was charged with "a felony violation of the CFAA." He was convicted and sentenced to 18 months of jail time but appealed the decision all the way to the Supreme Court. The Supreme Court overturned previous decisions and held that while Van Buren may have violated law enforcement policies, he was not in violation of the CFAA.

The Supreme Court found that questions of unauthorized access should be considered a "gates-up-or-down inquiry," meaning that a person either has access to a computer or certain information on a computer or they do not regardless of the context in which the information is being used. The alternative, as the majority opinion explained, would have disastrous implications for many computer users.[7]

The status of web-scraping and web-crawling under the CFAA is still ambiguous. More needs to be done to clarify the wording and enforcement of the act to avoid the dire consequences we saw in *U.S. v. Swartz.* This does not mean, however, that projects that use these methods should be avoided wholesale. As it stands, web-crawling and

---

[5] "NACDL - CFAA Cases," NACDL - National Association of Criminal Defense Lawyers, accessed August 3, 2021, https://www.nacdl.org/Content/CFAACases.
[6] Marcia Hofmann, "In the Wake of Aaron Swartz's Death, Let's Fix Draconian Computer Crime Law," Electronic Frontier Foundation, January 14, 2013, https://www.eff.org/deeplinks/2013/01/aaron-swartz-fix-draconian-computer-crime-law.
[7] Van Buren v United States (Supreme Court of the United States June 3, 2021).

web-scraping are typically fine as long as the information is public and measures are not taken to access restricted data or information.

Terms of Service

If a use violates a website's terms of use, it is best to avoid that use. A website's terms of use and terms of service pages are legally binding. While violating the terms of service is not a crime,[8] it may be a breach of contract and result in the removal of any content used from the site from your project. You should read the terms of service or terms of use of a web-page before using content from the site, including content gained from web-scraping or web-crawling.

In the case of APIs, there is typically robust documentation on how to use the API and what constitutes authorized usage. Because APIs are often developed to help users easily access information programmatically, the creators of the API are more likely to lay out how the API can be used. When looking at an API's documentation, look for pages marked "terms of use," "terms and conditions," or "developer guidelines" for more information. Some APIs will require you to sign up and agree to the terms and conditions before accessing the site's contents. If this is the case, failure to follow guidelines can result in a suspension of access to the API, as was the case with the website, Politiwoops.

Politiwoops is a website dedicated to hosting and displaying deleted tweets from politicians. The site was originally run by the Open State Foundation, a Dutch NGO, while the U.S. instance of the site was operated by the Sunlight Foundation. The U.S.

[8] Jamie Williams, "Ninth Circuit Doubles Down: Violating a Website's Terms of Service Is Not a Crime," Electronic Frontier Foundation, January 10, 2018, https://www.eff.org/deeplinks/2018/01/ninth-circuit-doubles-down-violating-websites-terms-service-not-crime.

installation ran from 2012 to 2015, when Twitter revoked access to its API[9] for failure to comply with the API's developer guidelines.

Politiwoops functions by monitoring tweets and, once they are flagged for removal by the user, archiving the tweet. Twitter's Developer Guidelines do not allow for the storage offline of deleted tweets and this resulted in the revocation of access to the API. Through a revision of the way the site worked and conversations with Twitter, Politiwoops went live again in 2016. Eventually, ProPublica took over for the Sunlight Foundation in running the site. The changes included only storing deleted tweets online and providing a list of tracked politicians and a way for politicians to opt out of being tracked.[10]

What we can learn from this case is twofold: ignoring terms of use or developer guidelines when using APIs can have real consequences for a project and working with an API outside of legal proceedings can result in a positive outcome for both parties.

Trespass to Chattel

Trespass to Chattel "refers to an act of intentional interference with the possessory rights of another's personal property."[11] In order for it to be enforceable in common law, there must be proof of damage. For example, if you were to use a web-crawler to crawl through multiple pages of a small business's website and that crawling affected how the website ran, this could be construed as trespass to chattel.

A prominent usage of trespass to chattel came in the court case, *EBay v Bidder's Edge.* Bidder's Edge, a website that aggregates items on different auction sites, used

---

[9]Jenn Topper, "Politwoops U.S. Is Back! : Sunlight Foundation," Sunlight Foundation, February 9, 2016, https://sunlightfoundation.com/2016/02/09/politwoops-u-s-is-back/.; Margarita Noriega, "Delete Your Tweets, Rewrite History? The Politwoops Controversy, Explained.," Vox, August 26, 2015, https://www.vox.com/explainers/2015/8/26/9211117/politwoops-delete-tweets..

[10] Derek Willis, "Politwoops," ProPublica, December 21, 2016, https://projects.propublica.org/politwoops/.

[11] Jonathan Band; Brandon Butler, "Overlapping Forms of Protection for Databases," in *Overlapping Intellectual Property Rights*, ed. Neil Wilkof and Shamnad Basheer (Oxford University Press, 2021), 27, https://libraopen.lib.virginia.edu/public_view/4x51hj12b.

web-crawling to get information from Ebay and other auction sites. Bidder's Edge originally had permission to scrape Ebay when a user entered in a search; however Bidder's Edge wanted to continuously crawl through Ebay's listings to make the site faster. Ebay issued a cease and desist to Bidder's Edge and the matter was eventually taken to courts as Ebay alleged financial damages.

The court ruled in Ebay's favor using the common law concept of trespass to chattel. This had the potential to negatively impact all forms of web-scraping and web-crawling, but "other courts have found that mere possessory interference is not sufficient harm for trespass to chattels liability. Rather, a showing of physical harm to the chattel or some obstruction of its basic function was necessary."[12]

Another example that has potential applications to trespass to chattel is the website Recipeasly. In 2021, Tom Redman announced he was creating a new website to "fix online recipes" by removing blog content and advertising and just leaving the recipes. This, presumably, would be done through web scraping. He experienced much backlash on this from social media and food recipe bloggers who worked hard to draft the content that was being deleted. Redmond's use was not a violation of U.S. copyright law because recipes are not copyrightable material. In fact, Redman was deleting the material that was copyrightable (the stories surrounding the recipes on blogs).[13]

Due to the backlash, the site was taken down hours after it was first released. While this never was taken to court, an argument could be made that Recipeasly caused financial harm by driving users away from websites with advertisements, meaning that the authors would not be able to profit from ad revenue. If the matter had been taken to court, it is possible that the concept of trespass to chattel could have been used.

Copyright

Even if your use is not prohibited by the CFAA or the terms of use, you may run into problems with copyright when using web-scraping, web-crawling, and APIs. Copyright

[12] Butler, "Overlapping Forms of Protection for Databases," 29.
[13] Jack Saxena, "What Is Recipeasly and Why Are Food Bloggers Mad About It?," Easter, March 1, 2021, https://www.eater.com/22307633/why-are-people-mad-at-recipeasly-recipe-blog-criticism.

protection in the United States is automatic and some content-creators may not know that their content requires their permission to be used. That's why it's important to familiarize yourself with copyright basics to understand if your use infringes someone's copyright.

The basics of copyright protection do not change for online content. User's rights such as fair use are still available and the rules governing what is in the public domain still apply. The additional considerations one must make regarding online content often fall into the categories described above, which offer additional protections and considerations.

An exception to this are databases (which can also take the form of an API). Because U.S. copyright law does not protect facts or ideas (rather they protect the original, creative *expression* of those ideas), copyright protection of databases can be a little tricky. Generally speaking, databases are not copyrightable because they do not exhibit a creative expression of the ideas or facts they hold. Content within a database, however, can be copyrightable (for example, in a database of song lyrics, the database itself may not be copyrightable, but the songs are). Exceptions can be made when the organization of a database is creative enough that it is an original work.[14]

Much of what we use for the basis of copyright applied to databases comes from the Supreme Court case, *Feist Publications, Inc. v, Rural Telephone Service Company, Inc.* Feist Publications, a publishing company that created phone directories, took phone numbers and their corresponding names/companies from Rural Telephone Service Co. (a telephone service company in Kansas). When Rural Telephone Service Co. sued, the court held that the arrangement of phone numbers did not pass the requisite "modicum of creativity" required in the law to be copyrightable. The court did, however, acknowledge that the published directory itself was copyrightable since it contained some original text.[15] This decision had notable effects on copyright law in the U.S. because it explicitly expressed that compilations of facts are not copyrightable unless

[14] Feist Publications, Incorporated v. Rural Telephone Service Company, Incorporated (Supreme Court March 27, 1991).
[15] Feist Publications, Incorporated v. Rural Telephone Service Company, Incorporated, 499.

they pass a threshold of originality or creativity. In the case of *Feist,* the directory was organized alphabetically which is not a creative nor original expression.[16]

Another example of how copyright is applied to online content comes from the more recent case, *HiQ Labs, Inc. v. LinkedIn Corp.* HiQ Labs is an analytics company that uses public data to generate reports on the state of employment. In generating these reports, HiQ Labs scraped publicly available data from Linkedin. In response, Linkedin issued a cease and desist to try to stop HiQ Labs from scraping the site. HiQ Labs, in turn, took the matter to court, claiming that the data was publicly available and Linkedin had no right to limit their access. In the summer of 2021, the matter was finally decided in favor of HiQ Labs.

The reasoning behind *HiQ Labs* relies on the public nature of Linkedin profiles as well as the fact that the content isn't created by Linkedin itself; rather, Linkedin hosts the information on its platform (meaning Linkedin does not hold a copyright for this content). While this is only one court case, the results are promising for those who want to use web-scraping of publicly available data in the future.[17]

[16] Feist Publications, Incorporated v. Rural Telephone Service Company, Incorporated, 499.
[17]Camille Fischer and Andrew Crocker, "Victory! Ruling in HiQ v. Linkedin Protects Scraping of Public Data," Electronic Frontier Foundation, September 10, 2019, https://www.eff.org/deeplinks/2019/09/victory-ruling-hiq-v-linkedin-protects-scraping-public-data.

Checklist General

- ☐ The terms of service or terms of use don't mention prohibiting the reproduction of site materials.
- ☐ The information is publicly available *(all below must be checked in order to check this box)*
  - ☐ No password or log-in is required to access the information
  - ☐ No sensitive information (addresses, medical statuses, etc.)
  - ☐ No technical barriers to downloading the material
- ☐ The information is legally obtained *(all below must be checked in order to check this box)*
  - ☐ There is no evidence suggesting that the information was made available without the owner's consent
  - ☐ There is no evidence that the information was obtained illegally or contains illegal material (pornographic, classified, etc).

**If you check all three of these boxes, continue to the next section that corresponds best with your use-case.*

Web-scraping and Web-crawling

- ☐ There is no robots.txt file that limits web-scraping or web-crawling
- ☐ If you are automatically scraping (crawling) pages, there are limitations set up so as not to overwhelm the server.

If you are posting the data:

- ☐ The data is protected by copyright
  - ☐ Your use fall into fair use
  - ☐ You have been given permission

If the data has is protected by copyright and you cannot check one of the above boxes, consider using a different source or reaching out to the copyright office for next steps

- ☐ Is there a chance that your posting of this data (even if it isn't copyrighted) will cause any damage to the owner (i.e. redirecting web-traffic and decreasing ad revenue).

  If you check the above box consider using a different source or reaching out to the copyright office for next steps

Databases

- ☐ The database isn't protected by copyright.
  - ☐ The information is **not** organized in a way that significantly adds to or changes the meaning of the information.
  - ☐ The database is in the public domain
- ☐ The information in the database is not protected by copyright

**if the database or the information in it is protected by copyright, look into whether your use could be considered fair use, reach out to the copyright owner, or contact the copyright office for assistance.*

Works Cited

Butler, Jonathan Band; Brandon. “Overlapping Forms of Protection for Databases.” In *Overlapping Intellectual Property Rights*, edited by Neil Wilkof and Shamnad Basheer. Oxford University Press, 2021. https://libraopen.lib.virginia.edu/public_view/4x51hj12b.

Crocker, Camille Fischer and Andrew. “Victory! Ruling in HiQ v. Linkedin Protects Scraping of Public Data.” Electronic Frontier Foundation, September 10, 2019. https://www.eff.org/deeplinks/2019/09/victory-ruling-hiq-v-linkedin-protects-scraping-public-data.

Feist Publications, Incorporated v. Rural Telephone Service Company, Incorporated (Supreme Court March 27, 1991).

Fischer, Camille, and Andrew Crocker. “Victory! Ruling in HiQ v. Linkedin Protects Scraping of Public Data.” Electronic Frontier Foundation, September 10, 2019. https://www.eff.org/deeplinks/2019/09/victory-ruling-hiq-v-linkedin-protects-scraping-public-data.

Gazarov, Petr. “What Is an API? In English, Please.” freeCodeCamp.org, December 19, 2019. https://www.freecodecamp.org/news/what-is-an-api-in-english-please-b880a3214a82/.

Hofmann, Marcia. “In the Wake of Aaron Swartz’s Death, Let’s Fix Draconian Computer Crime Law.” Electronic Frontier Foundation, January 14, 2013. https://www.eff.org/deeplinks/2013/01/aaron-swartz-fix-draconian-computer-crime-law.

NACDL - National Association of Criminal Defense Lawyers. “NACDL - CFAA Cases.” Accessed August 3, 2021. https://www.nacdl.org/Content/CFAACases.

NACDL - National Association of Criminal Defense Lawyers. “NACDL - Computer Fraud and Abuse Act (CFAA).” Accessed July 29, 2021. https://www.nacdl.org/Landing/ComputerFraudandAbuseAct.

Noriega, Margarita. “Delete Your Tweets, Rewrite History? The Politwoops Controversy, Explained.” Vox, August 26, 2015. https://www.vox.com/explainers/2015/8/26/9211117/politwoops-delete-tweets.

Perez, Martin. “Is Web Scraping Legal? Explanation and Examples | ParseHub.” ParseHub Blog, July 14, 2021. https://www.parsehub.com/blog/web-scraping-legal/.

———. “Web Scraping vs Web Crawling: What’s the Difference? | ParseHub.” ParseHub Blog, February 3, 2020. https://www.parsehub.com/blog/web-scraping-vs-web-crawling/.

Saxena, Jack. “What Is Recipeasly and Why Are Food Bloggers Mad About It?” Easter, March 1, 2021. https://www.eater.com/22307633/why-are-people-mad-at-recipeasly-recipe-blog-criticism.

Topper, Jenn. “Politwoops U.S. Is Back! : Sunlight Foundation.” Sunlight Foundation, February 9, 2016. https://sunlightfoundation.com/2016/02/09/politwoops-u-s-is-back/.

Van Buren v United States (Supreme Court of the United States June 3, 2021).

“What Is Recipeasly and Why Are Food Bloggers Mad About It? - Eater.” Accessed August 3, 2021. https://www.eater.com/22307633/why-are-people-mad-at-recipeasly-recipe-blog-criticism.

Williams, Jamie. “Ninth Circuit Doubles Down: Violating a Website’s Terms of Service Is Not a Crime.” Electronic Frontier Foundation, January 10, 2018. https://www.eff.org/deeplinks/2018/01/ninth-circuit-doubles-down-violating-websites-terms-service-not-crime.

Willis, Derek. “Politwoops.” ProPublica, December 21, 2016. https://projects.propublica.org/politwoops/.