INTELLIGENCE · OPEN-SOURCE INTELLIGENCE

"Web Scraping for Intelligence: Technique, Law and Ethics"

Web scraping collects information from websites automatically and at scale. This article explains how scraping works, what it is used for in business intelligence, the main legal questions in Canada and the United States, and the ethical practices that keep it responsible.

ModQu research article · 1,497 words · 16 references (APA 7th edition)

Much of the information that businesses want is published on websites: prices, product ranges, job postings, store locations, reviews, regulatory notices and news. Reading these pages by hand works for a few items, but not for thousands of products checked every day. **Web scraping** is the automated collection of information from web pages by software that requests pages, extracts the relevant parts and stores them in a structured form. It has become a standard tool of business and economic intelligence. It also raises real legal and ethical questions. This article explains the technique, its uses, the law as it stands in Canada and the United States, and practices that keep scraping responsible.

How scraping works

A scraper does what a browser does, but automatically. It requests a web page, receives the underlying code, locates the elements that contain the desired information, such as a product name and price, and saves them. A **crawler** follows links from page to page to discover what to collect. Olston and Najork (2010) describe the design of web crawlers, including the importance of politeness: limiting the rate of requests so as not to overload the sites being visited. Mitchell (2018) provides a practical guide to the techniques, from parsing page structure to handling pages that load content dynamically.

Many websites offer an **application programming interface (API)**, a structured way to request data, often with clear terms of use. Where one exists, it is usually more reliable and less legally risky than scraping the visible pages.

Uses in intelligence

Edelman (2012) describes how researchers use data collected from the internet to study prices, markets and behaviour that were previously hard to observe. Business and economic uses include:

- **Price monitoring.** Cavallo and Rigobon (2016) describe the Billion Prices Project, which collected daily online prices from hundreds of retailers in dozens of countries to measure inflation and study how prices change. Businesses use similar methods to track competitors' prices and promotions.
- **Market structure.** Chen et al. (2016) collected data from a large online marketplace to identify sellers using algorithmic pricing and to study how their prices behaved.

- **Product and assortment tracking.** Monitoring when competitors add or drop products, change specifications or alter stock levels.
- **Labour market signals.** Collecting job postings to see which skills competitors are hiring for and where.
- **Reviews and reputation.** Gathering reviews to analyse customer satisfaction across brands.
- **Regulatory and public notices.** Collecting announcements, tenders and decisions from government websites.

Landers et al. (2016) argue that scraping should be guided by theory: the analyst should decide in advance what the data are expected to represent and how they were generated, rather than collecting whatever is available. Salganik (2018) warns that data created for one purpose, such as a retailer's website, may be incomplete, change without notice and reflect the site's design more than the underlying reality.

The legal landscape

Scraping law is unsettled and varies between countries. What follows is general information, not legal advice; organisations should seek advice for their specific situation.

Unauthorised access

In the United States, much litigation has turned on the Computer Fraud and Abuse Act, which prohibits accessing a computer "without authorization" or exceeding authorised access. In Van Buren v. United States (2021), the Supreme Court read "exceeds authorized access" narrowly, focusing on whether a person enters parts of a system that are closed to them. In hiQ Labs, Inc. v. LinkedIn Corp. (2022), an appeals court held that scraping data from publicly accessible profiles was unlikely to amount to access "without authorization" under that law. The dispute did not end there, however: the litigation later turned to the platform's user agreement, a reminder that contract terms can bind a scraper even when anti-hacking law does not. Collecting data behind a log-in, or after being blocked, carries much greater risk.

Contract and terms of use

Websites typically set out terms of use, and many prohibit automated collection. Fiesler et al. (2020) examined the terms of service of a large number of social media platforms and found that most prohibited automated data collection, often in broad language. Whether such terms form a binding contract depends on how they are presented and accepted, but ignoring them is a legal and reputational risk.

Copyright and databases

Facts themselves are generally not protected by copyright. In Feist Publications, Inc. v. Rural Telephone Service Co. (1991), the United States Supreme Court held that the names and numbers in a telephone directory were not protected, because copyright requires originality. In Canada, the Copyright Act (1985) likewise protects original expression rather than facts, although an original selection or arrangement of material can be protected. Copying prices or product specifications is therefore different from copying articles, photographs or reviews in full. The European Union adds a separate database right, which can protect substantial investment in compiling data.

Privacy

Where scraped material includes personal information, privacy law applies. In Canada, the Personal Information Protection and Electronic Documents Act (2000) governs how private organisations collect, use and disclose personal information in commercial activity. Only narrowly defined categories of information are treated as publicly available for its purposes, so personal information posted online generally remains protected.

Krotov et al. (2020) review these legal issues and propose that organisations consider the explicit wishes of website owners, the potential for harm to the site, the use of personal or copyrighted material and the purpose of the collection.

Ethics

Legal permission is not the only test. Zimmer (2010) showed that collecting publicly visible personal information can still violate people's reasonable expectations about how it will be used. The Robots Exclusion Protocol, the "robots.txt" file that many sites publish, was formalised as a standard in 2022 (Koster et al., 2022). It tells automated agents which parts of a site the owner does not want crawled. It is not a security barrier and compliance is voluntary, but respecting it is a widely accepted norm.

Questions to answer before scraping

A short legal and ethical screen; any doubtful answer means stop and seek advice

Question	Why it matters
Is the data publicly available without logging in?	Bypassing log-ins or technical barriers raises legal risk
Do the site's terms of use allow automated collection?	Terms can form a contract that binds the scraper
Does robots.txt allow the pages you need?	The site's stated wishes; ignoring them is poor practice
Will collection put a noticeable load on the site?	Heavy traffic can harm the site and its users
Does the data include personal information?	Privacy law applies even to information posted publicly
Will you copy creative content, or only facts?	Facts are generally free to use; expression may be protected
Is there an official interface or data set instead?	An API or licence is usually safer and more reliable

Framework by ModQu, drawing on Krotov et al. (2020) and Fiesler et al. (2020)

ModQu.com

Figure 1. Questions to answer before scraping. Any doubtful answer is a reason to stop and seek advice. Framework by ModQu, drawing on Krotov et al. (2020) and Fiesler et al. (2020).

Data quality pitfalls

Even lawful, polite scraping can produce misleading intelligence if the data are not handled carefully. Common problems include:

- **Silent breakage.** Websites change their layout often. A scraper may continue to run but extract the wrong field, or nothing at all, without raising an error.
- **Personalised content.** Some sites show different prices or products depending on location, device or browsing history. A scraper sees one version, which may not be what customers see.
- **Incomplete coverage.** Products that are out of stock, hidden behind filters or shown only to logged-in customers may be missed, biasing averages.
- **Changing definitions.** A retailer may change how it lists sizes, bundles or promotional prices, creating apparent price changes that are artefacts.
- **Duplicates.** The same product may appear under several addresses or names.

Cavallo and Rigobon (2016) note that online prices are most useful when collected consistently over long periods, with careful matching of identical products over time. Comparing scraped figures with occasional manual checks, logging every change to the scraper and flagging sudden jumps for review are simple safeguards.

Good practice

Responsible scraping for business intelligence generally means:

1. **Prefer official sources.** Use APIs, data licences or published data sets where they exist.
2. **Collect only public pages.** Do not bypass log-ins, paywalls or technical blocks.
3. **Read and respect the terms.** If the terms prohibit scraping, seek permission or another source.
4. **Respect robots.txt.** Treat it as the owner's stated preference.
5. **Be polite.** Limit request rates, collect at quiet times and identify the scraper honestly.
6. **Minimise personal data.** Collect facts about products, prices and organisations rather than individuals.
7. **Store and document.** Record what was collected, when and from where, so results can be checked.
8. **Validate.** Check scraped data against manual samples, since page changes can silently break extraction.

An illustration

A Canadian online retailer of outdoor equipment wants to track competitors' prices on 2,000 products. Its team first checks whether competitors offer data feeds; none do. It reviews each competitor's terms of use and robots.txt file, excludes one site whose terms explicitly prohibit automated collection, and designs a scraper that visits the remaining sites once a day overnight, with pauses between requests, collecting only product names, prices and stock status. No personal information is gathered. A weekly manual check of a random sample confirms the extracted prices are accurate. When one competitor redesigns its site and the scraper begins returning blanks, the check catches the problem within days.

Conclusion

Web scraping lets organisations observe prices, products, hiring and public notices at a scale no manual process could match. It is a powerful intelligence technique, but one that sits in an unsettled legal area and touches on the interests of website owners and the people whose information appears online. Using official interfaces where possible, collecting only public facts, respecting terms and robots.txt, limiting load, avoiding personal data and validating results keep scraping both useful and responsible.

References

Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using online prices for measurement and research. *Journal of Economic Perspectives*, *30*(2), 151–178. https://doi.org/10.1257/jep.30.2.151

Chen, L., Mislove, A., & Wilson, C. (2016). An empirical analysis of algorithmic pricing on Amazon Marketplace. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 1339–1349). https://doi.org/10.1145/2872427.2883089

Copyright Act, R.S.C. 1985, c. C-42 (Canada).

Edelman, B. (2012). Using Internet data for economic research. *Journal of Economic Perspectives*, *26*(2), 189–206. https://doi.org/10.1257/jep.26.2.189

Feist Publications, Inc. v. Rural Telephone Service Co., 499 U.S. 340 (1991).

Fiesler, C., Beard, N., & Keegan, B. C. (2020). No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 187–196.

hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180 (9th Cir. 2022).

Koster, M., Illyes, G., Zeller, H., & Sassman, L. (2022). *Robots exclusion protocol* (RFC 9309). Internet Engineering Task Force. https://doi.org/10.17487/RFC9309

Krotov, V., Johnson, L., & Silva, L. (2020). Tutorial: Legality and ethics of web scraping. *Communications of the Association for Information Systems*, *47*, 539–563. https://doi.org/10.17705/1CAIS.04724

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological Methods*, *21*(4), 475–492. https://doi.org/10.1037/met0000081

Mitchell, R. (2018). *Web scraping with Python: Collecting more data from the modern web* (2nd ed.). O'Reilly Media.

Olston, C., & Najork, M. (2010). Web crawling. *Foundations and Trends in Information Retrieval*, *4*(3), 175–246. https://doi.org/10.1561/1500000017

Personal Information Protection and Electronic Documents Act, S.C. 2000, c. 5 (Canada).

Salganik, M. J. (2018). *Bit by bit: Social research in the digital age*. Princeton University Press.

Van Buren v. United States, 593 U.S. 374 (2021).

Zimmer, M. (2010). "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology, 12*(4), 313–325. https://doi.org/10.1007/s10676-010-9227-5