Business Horizons (2023) 66, 481—491

Available online at www.sciencedirect.com

ScienceDirect

KELLEY SCHOOL OF BUSINESS
INDIANA UNIVERSITY
www.journals.elsevier.com/business-horizons

# Big web data: Challenges related to data, technology, legality, and ethics

Vlad Krotov [a,*], Leigh Johnson [b]

[a] Murray State University, 625 Business Building, Murray, KY 42071, USA
[b] Murray State University, 132 Business Building, Murray, KY 42071, USA

**KEYWORDS**
Big data;
Web data;
Web scraping;
Law;
Ethics

**Abstract** The digital data available online is currently measured in zettabytes. These vast repositories of big web data are increasingly viewed as a strategic resource comparable in value to land, gold, and oil. This big web data can be extracted and analyzed by organizations to gain a better understanding of their internal and external environment and improve organizational performance. Because of these opportunities, automated retrieval and organization of web data (i.e., web scraping) for research projects is becoming a common practice. This article outlines the data-related, technical, legal, and ethical issues related to web scraping. Awareness of these issues can help researchers save time and resources and, most importantly, mitigate the potential risk of ethical controversies or lawsuits related to the retrieval and use of big web data.


## 1. The value of big web data

The increasing digitalization of various social processes has resulted in zettabytes (i.e., billions of gigabytes) of data deposits available on the World Wide Web (Cisco Systems, 2016). The vast data available on the web is mostly comprised of semistructured data in the form of web pages, online databases, emails, blog posts, photos, videos, etc. (Watson, 2014). The volume of digital big web data continues to grow due, in part, to the global COVID-19 crisis. Because of the pandemic, processes previously belonging to the physical, face-to-face realm (e.g., education, grocery shopping) are increasingly digitalized and conducted online—either willingly or due to government mandates (Brough & Martin, 2021). This contributes to the growth of traffic and new digital record creation on the web.

These vast repositories of web data are increasingly viewed as the most valuable resource of the 21st century—similar to oil in the 20th century (Alharthi et al., 2017; The Economist, 2017). When people use the web either actively (e.g., via

* Corresponding author
*E-mail addresses:* vkrotov@murraystate.edu (V. Krotov), ljohnson1@murraystate.edu (L. Johnson)

https://doi.org/10.1016/j.bushor.2022.10.001
0007-6813/

blog or social posts) or passively (e.g., via cookies or search engine queries), they leave digital footprints that are rich in psychological and social meaning (Speckmann, 2021). These digital footprints can be used to gain an intricate understanding of numerous processes, behaviors, relationships, and interactions in the real world at individual, group, organizational, national, and even global levels (Dabirian et al., 2017). This understanding can be valuable for gaining a better perspective of customers, improving current business processes, supporting decision-making, informing organizational strategies, or creating data products that either enhance current products and services or create new value propositions for customers (Grover et al., 2018).

For example, the Center for Computer and Information Technology (CIT) at Murray State University, a midsize, regional university in western Kentucky, scrapes public data via LinkedIn to identify and profile all university alumni who graduated from current and former programs related to computing. This builds a closer connection with Murray State's alumni base to increase their engagement in decision-making processes.

In addition, Starbucks is known for proactively sourcing social media posts that are related to the company and its coffee brands. This data provides insight into how consumers think about the company and ways to change certain aspects of its operations. Some digital experts have labeled such practices as a social media listening strategy.

Web data is now heavily utilized in sports as well. Via web scraping, coaches, managers, and fantasy players use athlete data to conduct performance analytics, which aids in the selection of appropriate players for a team—real or make-believe. Advertisers use market value analytics to determine a player's popularity and social value.

Unfortunately, web data is often sourced without considering the legal and ethical implications of sourcing and analyzing it. In this article, we discuss the nature of big web data as well as the numerous sociotechnical issues that must be addressed to make sourcing and analyzing big web data not only technologically possible but also compliant with the existing legislature and ethical norms.

## 2. Web scraping: A sociotechnical perspective

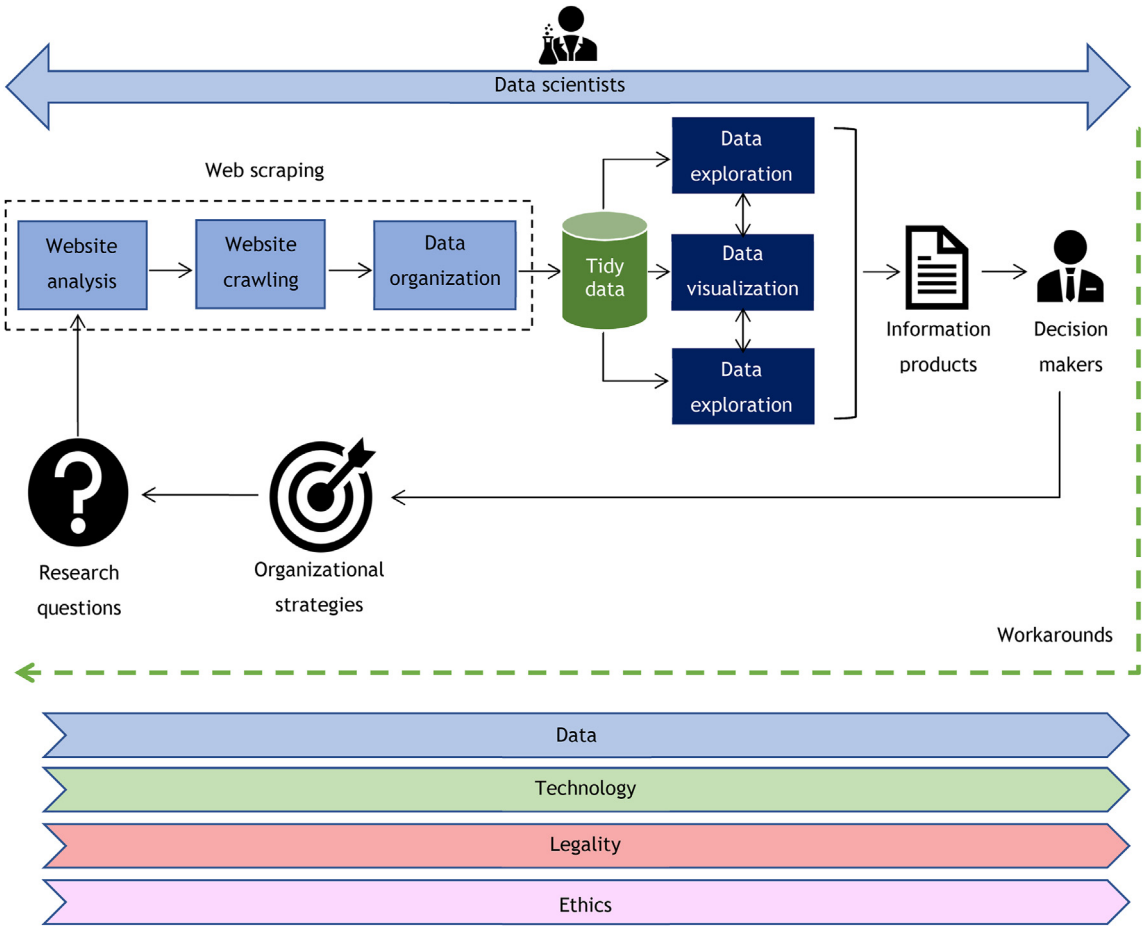
Before value is extracted from web data, it must be harvested via *web scraping* (i.e., automatically extracting and organizing web data that enables further analysis; Krotov & Tennyson, 2018, 2021). However, web scraping is different from simply downloading data from an online source using a file or an application programming interface (API; Speckmann, 2021). Rather, web scraping uses tools—either ready-made or custom—to retrieve and organize unstructured or semistructured data from web repositories that do not immediately provide easy and approved access to their data (e.g., via files or APIs).

Web scraping is conducted as a part of a broader organizational data science process (see Figure 1). We view both web scraping and data science as part of a complex, sociotechnical work system (Alter, 2013), which includes data-related activities, components, and participants that are supported by technological tools. This system is then carried out within a broader social environment subject to ethical and legal constraints.

The data science process starts with a research question that relates to the internal or external environment of an organization. These research questions regarding organizational strategies aim to create value for customers and improve organizational performance (Provost & Fawcett, 2013). Once a research question is formulated, the necessary data is sourced from the web via website analysis, website crawling, and data organization. The result of these activities is tidy data (i.e., a well-organized dataset that is stored securely for further analysis; Wickham, 2014). This dataset is further explored, visualized, and analyzed using modern data analysis tools to produce informational products (e.g., reports and dashboards) that can be used by organizational decision-makers to formulate or finetune organizational strategies. These data-related activities are driven by data and supported by the organization's existing technologies.

The data science process is supervised by data scientists who cooperate with IT workers and business users within their organization. Sourcing messy and diverse data from the web to create new or enhance existing data-driven products or services requires numerous goal-oriented workarounds. These workarounds modify existing technology to overcome data-related, technological, legal, and ethical challenges, as well as the changes in the internal and external environment of the organization (Alter, 2014). All these activities and workarounds regarding big web data sourcing and analysis are carried out within the broader social environment, which is comprised of stakeholders whose rights and well-being are protected by the law and ethical norms.

**Figure 1. Web scraping as a part of the data science process**




Current literature on web scraping often overlooks the ethical and legal aspects related to sourcing, analyzing, and using big web data and is typically comprised of two dominant streams. The first stream discusses the benefits of using big web data in organizations and academic research projects, whereas the second stream discusses numerous technical issues and workarounds related to handling big web data. The subtler, more significant ethical and legal aspects of web data retrieval are rarely discussed (Krotov et al., 2020). Ignoring the ethical and legal issues that relate to big web data can lead to serious ethical controversies or lawsuits (Ives & Krotov, 2006).

Thus, the main goal of this article is to educate researchers not only about the nature of big web data and the technical aspects of handling it but also about legal and ethical dilemmas related to web data retrieval. We believe that broad awareness of sociotechnical issues surrounding big web data can help researchers decrease the time required to source and organize data, minimize the resources necessary for data collection, and avoid ethical controversies or even lawsuits in some extreme cases.

## 3. Data-related challenges of web scraping

In this article, we define big web data as any large repository of digital data available on the World Wide Web (i.e., the web). This data can come in different forms: online databases, emails, blog posts, photos, videos, etc. (Watson, 2014). Big web data can also be both quantitative and qualitative and can be structured, semistructured, or unstructured.

Big web data is inherently different from the traditional data sources utilized by industry and academic researchers in the past. Besides its growing strategic value for individuals and organizations—as articulated in Section 1—it is also

characterized by vast volume, variety, velocity, and veracity (Wamba et al., 2015). These last four dimensions of big web data and their related challenges are further discussed in Section 3.

### 3.1. Volume

The current volume of data available online is staggering, as it is measured in zettabytes (Cisco Systems, 2016). Not only does this data surpass the ability of most individuals to wrap their brains around its vastness and complexity, but it also pushes existing technologies and methods for sourcing, storing, and analyzing data to their limits.

Popular spreadsheet software (e.g., Excel) and relational databases (e.g., Microsoft Access) are hardly capable of accommodating datasets that contain billions of rows of data. For example, Microsoft Excel has a limit of one million rows of data, yet this limitation may have caused the loss of nearly 16,000 COVID-19 test results by Public Health England (Hern, 2020). Vast, complex, and often messy big web data can hardly be organized using one of the "normal forms" within relational databases, such as Microsoft Access.

Thus, working with these vast volumes of data requires new tools and platforms. For example, NoSQL databases—such as Apache Cassandra—are less restrictive regarding the format for data storage, making them more compatible with the inherent messiness of big web data. In addition, many data science projects involving big web data often push the limits of existing IT infrastructure, but big web data analysis can be accomplished using tools installed on a cloud platform, such as Amazon Web Services (AWS). There, computing resources can be rapidly scaled to accommodate the necessary volume of data and resulting computational complexity.

Even if the necessary technical solutions for storing massive volumes of web data are implemented, organizations face the cognitive limitations of organizational decision-makers, such as managers, senior executives, board members, and external stakeholders (Merendino et al., 2018). These individuals and groups within organizations may simply lack the necessary cognitive capacity, mental models, and behaviors for utilizing the massive volumes of data they are responsible for. Thus, accommodating large quantities of big web data requires technical changes, employee training, and modifications to existing organizational decision-making processes.

### 3.2. Variety

Much of the data available on the web is comprised of semistructured, qualitative data in the form of web pages, blog posts, social media posts, customer reviews, etc. (Watson, 2014). Compared to quantitative data, qualitative data is more difficult to structure and organize. Typically, web data repositories that contain qualitative data are built using a wide variety of technologies and standards. For example, the data on web pages can be organized with the help of markup languages such as HTML and CSS. Web application data can be transmitted using XML or JSON. In addition, some financial institutions use XBRL (i.e., an XML extension designed specifically for posting and transmitting financial reporting data over the web). However, these technologies are problematic because they are not as strict as programming languages. Rather, they are more akin to technical recommendations for how web data should be organized. These recommendations are subject to different interpretations and can be implemented differently at the technical level, often resulting in a lack of compatibility between web repositories and web scraping tools. For example, an automatic web scraping tool may work on one website but then require modifications or customization to retrieve data from another website—due, in part, to how markup languages organize data on the latter website.

### 3.3. Velocity

Web data is generated and updated with high velocity. This creates numerous challenges for those wishing to harness data in a constant state of flux. Harvesting data from a web repository may require hours, if not days, of continuous work by a computer or an information system comprised of many computing nodes. By the time a web scraping task is complete, the harvested data may already be outdated and contain incomplete or incorrect information. For example, it may take many hours to scrape tens of thousands of job ads from an employment website (e.g., Dice.com or Indeed.com). As a web scraping script is being executed, some job postings are added, removed, or modified by the recruiters and employers using these websites. Thus, by the time the web scraping task is completed, the data obtained is already somewhat inaccurate and outdated.

The velocity with which data and its underlying structure change also hinder the development of

reliable web scraping tools. Most likely, a newly developed web scraping tool will work only temporarily. To continue functioning, it must be updated to reflect the changes in data volume, underlying structure, and storage technology. As such, developing automated tools for web scraping becomes a never-ending race—like how web data never stops growing and morphing.

### 3.4. Veracity

Another important characteristic of big web data is veracity (Wamba et al., 2015). The internet was created to enable open, voluntary, and anonymous interactions between people and organizations. Because of this, there is an inherent uncertainty associated with the availability, reliability, and validity of web data. A web data repository available today can be removed tomorrow. Even if data is permanently available on the web, issues related to its reliability and validity are always a big concern.

The data contained in online reviews for various products and services provides a good illustration of this problem. While serving as a useful source of information for consumer opinion and sentiment regarding products and services offered online, the quality of this data is often plagued by fraudulent reviews. For example, some vendors post fake positive reviews for their products and services in hopes of boosting sales. On the other side of the spectrum, dissatisfied customers may post exaggerated negative reviews based on their experiences with the same products and services. Companies can also post negative reviews to damage competitors' reputations. However, companies can sometimes remove reviews from their websites if they are believed to be fake or suspicious. Some online platforms allow vendors to dispute reviews they believe are incorrect or fake, yet those gathering and analyzing these online reviews can never be completely sure of their validity.

## 4. Technical challenges of web scraping

Given the issues related to the volume, variety, velocity, and veracity of big web data, retrieving and organizing such data can rarely be done by individual researchers or large teams of data analysts (Krotov & Tennyson, 2018). Instead, researchers resort to various web scraping tools and technologies to automate some aspects of big web data collection and organization (Krotov et al., 2020; Krotov & Tennyson, 2021).

### 4.1. Point-and-click web scraping tools

In recent years, numerous point-and-click web scraping tools have emerged (Krotov et al., 2020). These software tools allow one to scrape data from the web with minimal knowledge of programming and web technologies. Some of these tools are stand-alone software applications (e.g., OutWit Hub), while others are cloud-based platforms (e.g., import.io). In addition, some web scraping tools contain elements of artificial intelligence (AI) or allow coding to create some degree of customization or modification of the tool. Most applications have an intuitive, graphical user interface (GUI) that simplifies an individual's interactions with these tools.

However, these visual, point-and-click tools may not properly collect data for a research project. Web data is organized using a wide variety of technologies, conventions, and proprietary solutions, and web scraping tools are not smart enough to consistently and reliably obtain the exact data elements the researcher needs. Sometimes, a web scraping tool grabs a data element in a way that requires further data processing (e.g., by including several data elements as one). Other times, the tool simply does not work properly due to changes in the website's structure and data elements. To make things worse, some websites resist scraping from well-known web scraping tools by requiring users to authenticate themselves before gathering data or by restricting all robots from crawling the website and downloading data. While some developers of web scraping tools are starting to include features that address these issues (e.g., automatic website authentication or AI elements that make a web robot appear human to a website), they are not always ahead of the website developers who make their sites inaccessible to web robots—either intentionally or unintentionally.

### 4.2. Custom web scraping tools

Because point-and-click tools do not always work, some researchers choose to develop custom tools specific to their research projects using popular programming languages such as Python or R. Performing web scraping with personalized tools usually requires undergoing the following overlapping phases: website analysis, website crawling, and data organization (see Figure 1).

Website analysis requires examining the underlying code of a website to understand how the needed web data is stored at the technical level.

This analysis often requires some knowledge of the World Wide Web client-server architecture, various markup languages used for representing data on the web (e.g., HTML, CSS, XML, XBRL), and popular web databases (e.g., MySQL). Some websites provide application programming interfaces (APIs) with related manuals to make accessing data easier. These APIs can be used by developers to access data directly from the website's databases—eliminating the necessity to scrape data from public web pages. These APIs allow researchers to develop custom web scraping tools using less time and effort.

*Website crawling* is the automatic browsing of a website to retrieve necessary data. The browsing process is automated by developing and running a script written using R or Python (i.e., popular programming languages/data analysis tools in the data science community). Moreover, these languages contain libraries (e.g., the "rvest" package in R or the Beautiful Soup library in Python) with ready-made functions for automatic web crawling and data wrangling.

After a research project's necessary web data is retrieved from a web repository, it must be cleaned and organized via various technological tools for further analysis. Given the sheer volume of big web data, its organization is usually automated via an R or Python script. These popular programming languages contain numerous data manipulation libraries that are quite useful for cleaning and organizing web data.

The three intertwined phases of web scraping usually require some degree of human involvement or supervision. For example, a script developed for web scraping can halt its execution halfway through the web scraping task due to a power outage in the building or a network error. If this happens, a researcher collecting this data must analyze the error and restart the script, often with various modifications. Thus, web scraping—with the help of custom tools—is hard to fully automate.

## 5. Legal challenges of web scraping

While advancements in web scraping tools and technologies make web data easier to access for a variety of users, the legality of web scraping activities is complicated and often overlooked (Snell & Menaldo, 2016). No current legislation directly addresses web scraping. Rather, it is guided by a myriad of federal and state statutes and common law theories, such as illegal access and use of data, breach of contract, copyright infringement, trespass to chattels, and trademark misappropriation (Dreyer & Stockton, 2013; Snell & Menaldo, 2016). These legal elements, their applicability to web scraping, and a summary of relevant cases (see Table 1) are discussed throughout Section 5.

### 5.1. Illegal access and use of data

The Computer Fraud and Abuse Act (CFAA) and equivalent laws serve as the basis for most legal disputes involving web scraping. The CFAA prohibits the intentional, unauthorized access of a computer or access to a computer that exceeds the authorization given to a user (Computer Fraud

**Table 1. Court cases addressing web scraping**

| Theory | Relevant cases |
|---|---|
| **Illegal access** | EF Cultural Travel BV v. Zefer Corp., 318 F.3d 58 (1st Cir. 2003); Southwest Airlines Co. v. Farechase, Inc. 318 F. Supp. 2d 435 (N.D. Tex. Mar. 19, 2004); Craigslist Inc. v. 3Taps Inc., 942 F.Supp 2d 962 (N.D. Cal. 2013); Facebook, Inc. v. Power Ventures, Inc. et al., 844 F.3d 1058 (9th Cir. 2016); EarthCam, Inc. v. OxBlue, 703 Fed. Appx. 803 (11th Cir. 2017); Alan Ross Machinery Corp. v. Machinio Corp, 2018 WL 6018603 (N.D. Ill. Nov. 16, 2018); Ticketmaster L.L.C. v. Prestige Entertainment, Inc. et al., No. 17-cv-07232, 2018 WL 654410 (C.D. Cal. Jan. 31, 2018); hiQ Labs, Inc. v. LinkedIn Corp, 938 F.3d 985 (9th Cir. 2019); LinkedIn Corp. v. hiQ Labs, Inc., No. 19-1116, 593 US (GVR Order June 14, 2021) |
| **Breach of contract** | Alan Ross Machinery Corp. v. Machinio Corp, 2018 WL 6018603 (N.D. Ill. Nov. 16, 2018); Facebook, Inc. v. Power Ventures, Inc. et al., 844 F.3d 1058 (9th Cir. 2016) |
| **Copyright** | Kelly v. Arriba Soft Corp., 336 F.3d 811 (9th Cir. 2003); Associated Press v. Meltwater US Holdings, Inc. 931 F. Supp. 2d 537 (S.D.NY 2013) |
| **Trespass to chattels** | eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058 (N.D. Cal. 2000); Intel Corp. v. Hamidi, 71 P. 3d 296 (2003) |
| **Trade secrets** | Compulife Software Inc. v. Newman, 959 F.3d 1288 (11th Cir. 2020) |

and Abuse Act, 1986). Most legal opinions involving web scraping focus on what constitutes unauthorized access under the CFAA. Although these court decisions are not always consistent (Sellars, 2018), they provide some guidance on how the CFAA can be applied to web scraping activities.

Courts initially considered whether a website's terms of use or terms of service policy prohibited web scraping activities, noting that access could be unauthorized and illegal when the website user takes some type of affirmative action to become a party to the terms of use or terms of service. Other courts have held that violation of the user agreement alone does not constitute illegal access.

Courts are divided regarding whether a cease-and-desist order prevents web scraping activities and makes access to the website unauthorized under the CFAA. In 2019, the Ninth Circuit Court of Appeals allowed web scraping of a website's public data, despite a cease-and-desist order and an IP block. Under the Supreme Court's directive, the Ninth Circuit is reconsidering whether technical measures, such as a cease-and-desist order, constitute unauthorized access (a "gates down" approach) or whether publicly available data is always accessible (a "gates up" approach).

### 5.2. Breach of contract

Website owners may prohibit programmatic access to a site by preventing web scraping activities in its terms of use or terms of service policies. Violating such policies may lead to a breach of contract claim by a website owner (Dreyer & Stockton, 2013) if website users explicitly agreed to comply with such policies (e.g., by clicking on a checkbox) and caused material damage to the website.

### 5.3. Copyrighted material

Web scraping activities can lead to copyright infringement claims—especially when scraped copyrighted data is used for financial gain (Dreyer & Stockton, 2013). Even open, public data may be subject to some legal restrictions (e.g., no commercial use) outlined in the user agreement. While copyright law does not prohibit the act of web scraping, it can prevent the republishing of data or copyrighted information. Although the fair use principle allows copyrighted material to be published in a new or original way, courts have reached different conclusions regarding whether publishing scraped materials results in fair use, especially when the data is available for purchase.

### 5.4. Trespass to chattels

Overloading or damaging a website by web scraping could lead to a trespass to chattels claim (Dreyer & Stockton, 2013). This can happen when a web scraping project prevents others from using a web server (e.g., due to overloading its capacity), but the damage must be material for the website owner to receive financial compensation. Since this damage is often hard to prove, a trespass to chattels claim is not common in web scraping cases (Gold & Latonero, 2018).

### 5.5. Trade secrets

In certain circumstances, scraping of information—even publicly available data—may constitute trade secret misappropriation. In May 2020, the Eleventh Circuit Court of Appeals ruled that when bots used improper means to recreate a portion of an insurance database by scraping considerable amounts of proprietary information, such web scraping was an unlawful misappropriation of a trade secret under state law. Although the information was public to the extent that individuals could obtain an insurance quote from the website, scraping an infeasible number of quotes could be improper, as long as the website took steps to maintain the database's secrecy.

## 6. Ethical challenges of web scraping

While court precedents provide some guidance on the legality of web scraping, the ethics of such activities have only recently been addressed. In this article, we narrowly define ethics as "a set of concepts and principles that guide us in determining what behavior helps or harms sentient creatures" (Paul & Elder, 2006). Unscrupulous web scraping can harm the sentient creatures associated with a particular website, including website designers, owners, customers, or visitors. Issues caused by web scraping that concern privacy, trade secrets, organizational value, discrimination, bias, and data quality are described in Section 6.

### 6.1. Individual privacy and rights of research subjects

Web scraping projects can compromise an individual's privacy when the individual interacts with available activities via a website (Mason, 1986). Even an anonymous website user can be identified by matching collected data with other

**Table 2. Legal, ethical, data, and technical dimensions of web scraping**

| | |
|---|---|
| Legality | • Is the website's data private (i.e., protected from public access)?<br>• Does the website contain copyrighted data?<br>• Does the website's terms of use policy prohibit web scraping?<br>• Is the data being scraped available for purchase?<br>• Does the project for which the data is collected involve illegal or fraudulent use of the data?<br>• Does the use of data constitute misappropriation of a trade secret?<br>• Can web scraping cause material damage to the web server hosting the website (e.g., by preventing others from accessing the website)?<br>• Has the website's administrator sent a cease-and-desist order, blocked your IP address, or restricted website access in some other way? |
| Ethics | • Can using the data obtained from the website potentially compromise individual rights to privacy or nondiscrimination?<br>• Can such data reveal confidential information about various organizations affiliated with the website?<br>• Can the project for which data is scraped produce deliverables that compete with the existing products or services offered by the website?<br>• Can the data lead to ill-informed decision-making with significant negative consequences for various stakeholders? |
| Data | • Does the data contain valid, reliable information?<br>• Is the website's data available in suitable format?<br>• Will the website's data be accessible long enough to complete the research project? |
| Technology | • Do we have appropriate tools for automatic collection and organization of the website's data?<br>• Can our existing computational infrastructure support collection, organization, analysis, and storage of big web data? |

Sources: Krotov & Silva (2018), Krotov et al. (2020)

sources (Ives & Krotov, 2006). For example, when the now-defunct AOL search engine released 500,000 anonymized search engine queries to the public, many behind these searches were quickly identified by matching the information obtained from their searches (e.g., names, geographical locations,) with data available in other web repositories. Even when individual privacy is not compromised, a website user may not have a customer's permission to access consumer data from a website, and the use of such data without consent violates the rights of research subjects (Buchanan, 2017). In light of recent privacy scandals involving companies such as Facebook, Marriot Hotels, and Panera Bread, such violations can have serious consequences for website owners.

## 6.2. Organizational privacy and trade secrets

It can be argued that organizations, just like individuals, have a right to maintain confidentiality regarding their assets and operations (Mason, 1986). Web scraping activities can potentially reveal trade secrets, website ownership, or other proprietary information. For example, by automatically counting employment ads on an online recruitment website, the website's market share and revenues can be approximated. Web scraping can also reveal flaws in the way the data is stored by the website, opening up the possibility of a lawsuit against the website owner (Ives & Krotov, 2006). Such activities can harm the reputation of the organization behind a website, causing substantial financial loss.

## 6.3. Diminishing value for the organization

The profitability of many websites is often tied to advertisement exposure or the sale of their data products. Web scraping tools can be used to access data directly from the website, bypassing the website's main web interface that contains ads for website visitors. If one uses tools to forgo the web

interface made for humans, the website owners lose the ability to monetize their website content. Moreover, a web scraping project can result in a deliverable (e.g., a report) that, without infringing on the copyright, disincentivizes a potential client from purchasing data products or services (Hirschey, 2014). For example, an employment website can offer numerous proprietary products concerning employee recruitment and employment trends. If a web scraping tool allows someone to obtain this data for free, potential clients will be less likely to purchase these proprietary products, causing financial losses for the website owners.

### 6.4. Discrimination and bias

Bias can affect data and algorithms and influence discriminatory decision-making. For example, facial recognition software has incorrectly identified African American men at a higher rate than other groups, leading to false arrests. Inherent bias in search engine technology and other tools can reflect underlying stereotypes and will not generate useful information. Using biased data in hiring and credit decisions can have a discriminatory impact on women and certain minority groups (Hassan & Gezahegn, 2020). Thus, organizations must ensure that web data used for analysis does not harbor such bias.

### 6.5. Data quality and decision-making

Web data may be inaccurate or incomplete due to its veracity, leading decision-makers to draw incorrect conclusions from data collected via web scraping activities (Clarke, 2016). However, markets often encourage web scraping activities that collect and sell low-quality data (Martin, 2015). These errors in web data can flow from one user to the next, and without the ability to modify or update inaccuracies, false data is perpetuated by a value chain (Someh et al., 2019). This can contribute to poor decision-making, financial losses, and negative impacts on consumers and other stakeholders (Wigan & Clarke, 2013).

## 7. Recommendations for web scraping

Data available online may appear to be abundant, free, and easily obtainable to address important research questions or improve organizational performance. Unfortunately, this can be a false perception. As explained in this article, harnessing big web data requires a good understanding of (1) the nature of the data itself, (2) various tools and technologies used for harvesting, organizing, and storing data, and (3) legal and ethical issues surrounding the collection and use of big web data. By far, the legal and ethical issues are the most important and abundant (Krotov et al., 2020). Before commencing a web data collection project, one should reflect on the legal and ethical questions related to the big web data of interest (see Table 2).

If a potential legal or ethical issue surrounding big web data is detected, this does not mean that the research project should be halted. Numerous workarounds can be implemented to mitigate various ethical and legal problems. For example, copyrighted data can still be used per the fair use principle. Permission to collect copyrighted data or to automatically crawl a website protected by robots.txt can be obtained from the website's owner. Researchers can also take the necessary precautions to protect the privacy of those whose data is collected by anonymizing it and storing it in a secure, encrypted storage location. In any case, a potential ethical or legal issue requires thorough reflection regarding how such controversies can be avoided.

After the legal and ethical issues related to a web scraping project are fully addressed, researchers can start looking closely at the data itself and the technical tools available for scraping it from the web. If data contains valid, reliable information that is available in a format suitable for automatic data collection (e.g., HTML, XML, JSON), researchers can consider existing web scraping applications to harvest this data. For starters, one should look at visual, cloud-based tools, such as import.io. In many cases, researchers only need one of these tools to obtain the necessary web data.

If none of the tools work as intended, the researchers should consider developing custom tools for web data collection using programming languages such as R or Python. This requires a good understanding of various web formats and technologies, such as HTML, CSS, XML, and XBRL. Often, retrieving and organizing web data requires an understanding of databases. Most importantly, one must be comfortable with programming and have hands-on knowledge of programming languages. This is necessary as such programs often develop tools for web crawling and scraping due to the availability of libraries that implement various interactions with a website.

## 8. Legality and ethics in data science

As previously argued, the issues related to legality and ethics span not only the typical phases of web scraping but also the entire data science work system (see Figure 1). Today, researchers increasingly rely on artificial intelligence (e.g., machine learning, sentiment analysis, and/or deep learning) to extract meaning and interference from web data (Lee et al., 2020). The notion that AI is inherently ethical, objective, and intelligent has been challenged in recent years (White & Lidskog, 2021). For example, AI can be used to triangulate an individual by combining data from various data sources and discovering patterns and facts about that individual that violate their right to privacy. AI also has been accused of generating machine learning modules that are biased toward certain societal groups (Hassan & Gezahegn, 2020). Overall, humans utilizing AI are not fully aware of its nature and limits, as well as the possible risks that this lack of understanding can create (White & Lidskog, 2021).

Yet, scientists and policymakers are arguably too focused on the technical aspects of big data analysis that use AI and rarely devote enough thought to the legal, ethical, and broader social implications of these new research approaches (White & Lidskog, 2021). Thus, adherence to the highest standards of legality and ethics when collecting and organizing big web data is necessary. However, these efforts do not inherently ensure the legality and ethics of the overall data science project, nor protect individuals, organizations, and broader society from harm. After big web data is gathered and organized, these issues should be heavily considered and dealt with throughout the entire research project.

## References


Alharthi, A., Krotov, V., & Bowman, M. (2017). Addressing barriers to big data. *Business Horizons, 60*(3), 285—292.

Alter, S. (2013). Work system theory: Overview of core concepts, extensions, and challenges for the future. *Journal of the Association for Information Systems, 14*(2), 72—121.

Alter, S. (2014). Theory of workarounds. *Communications of the Association for Information Systems, 34,* Article 55.

Brough, A. R., & Martin, K. D. (2021). Consumer privacy during (and after) the COVID-19 pandemic. *Journal of Public Policy and Marketing, 40*(1), 108—110.

Buchanan, E. (2017). Internet research ethics: Twenty years later. In M. Zimmer & K. Kinder-Kurlanda (Eds.), *Internet research ethics for the social age: New challenges, cases, and contexts* (pp. xxix—xxxiii). Bern, Switzerland: Peter Lang International Academic Publishers.

Cisco Systems. (2016). *Cisco visual networking index: Forecast and methodology, 2014—2019* [White Paper]. Available at https://www.brodeur.com/wp-content/uploads/2016/01/white_paper_c11-481360.pdf

Clarke, R. (2016). Big data, big risks. *Information Systems Journal, 26*(1), 77—90.

Computer Fraud and Abuse Act, 18 U.S.C. § 1030 (1986)

Dabirian, A., Kietzmann, J., & Diba, H. (2017). A great place to work!? Understanding crowdsourced employer branding. *Business Horizons, 60*(2), 197—205.

Dreyer, A. J., & Stockton, J. (2013, July 15). Internet 'data scraping': A primer for counseling clients. *New York Law Journal*. Available at https://www.law.com/newyorklawjournal/almID/1202610687621

Gold, Z., & Latonero, M. (2018). Robots welcome? Ethical and legal considerations for web crawling and scraping. *Washington Journal of Law, Technology, and Arts, 13*(3), 275—312.

Grover, V., Chiang, R. H., Liang, T. P., & Zhang, D. (2018). Creating strategic business value from big data analytics: A research framework. *Journal of Management Information Systems, 35*(2), 388—423.

Hassan, F., & Gezahegn, H. (2020, August 9). Addressing racial bias in AI: A guide for curious minds. *Towards Data Science*. Available at https://towardsdatascience.com/addressing-racial-bias-in-ai-a-guide-for-curious-minds-ebdf403696e3

Hern, A. (2020, October 6). Covid: How Excel may have caused loss of 16,000 test results in England. *The Guardian*. Available at https://www.theguardian.com/politics/2020/oct/05/how-excel-may-have-caused-loss-of-16000-covid-tests-in-england

Hirschey, J. K. (2014). Symbiotic relationships: Pragmatic acceptance of data scraping. *Berkeley Technology Law Journal, 29*(4), 897—927.

Ives, B., & Krotov, V. (2006). Anything you search can be used against you in a court of law: Data mining in search archives. *Communications of the Association for Information Systems, 18,* Article 29.

Krotov, V., Johnson, L., & Silva, L. (2020). Tutorial: Legality and ethics of web scraping. *Communications of the Association for Information Systems, 47*(22), 555—581.

Krotov, V., & Silva, L. (2018). Legality and ethics of web scraping. In *Twenty-fourth Americas Conference on Information Systems*. Available at https://www.researchgate.net/publication/324907302_Legality_and_Ethics_of_Web_Scraping

Krotov, V., & Tennyson, M. (2018). Scraping financial data from the web using the R language. *Journal of Emerging Technologies in Accounting, 15*(1), 169—181.

Krotov, V., & Tennyson, M. (2021). Web scraping in the R language: A tutorial. *Journal of the Midwest Association for Information Systems, 2021*(1), Article 5.

Lee, L. W., Dabirian, A., McCarthy, I. P., & Kietzmann, J. (2020). Making sense of text: Artificial intelligence-enabled content analysis. *European Journal of Marketing, 54*(3), 615—644.

Martin, K. (2015). Ethical issues in the big data industry. *MIS Quarterly Executive, 14*(2), 67—85.

Mason, R. O. (1986). Four ethical issues of the information age. *MIS Quarterly, 10*(1), 5—12.

Merendino, A., Dibb, S., Meadows, M., Quinn, L., Wilson, D., Simkin, L., & Canhoto, A. (2018). Big data, big decisions: The impact of big data on board level decision-making. *Journal of Business Research, 93*(C), 67—78.

Paul, R., & Elder, L. (2006). *The thinker's guide to understanding the foundations of ethical reasoning*. Santa Barbara, CA: Foundation for Critical Thinking.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data, 1*(1), 51—59.

Sellars, A. (2018). Twenty years of web scraping and the computer fraud and abuse act. *Boston University Journal of Science and Technology, 24*(2), 372—415.

Snell, J., & Menaldo, N. (2016, June 1). Web scraping in an era of big data 2.0. *Bloomberg Law*. Available at https://news.bloomberglaw.com/tech-and-telecom-law/web-scraping-in-an-era-of-big-data-20

Someh, I., Davern, M., Breidbach, C., & Shanks, G. (2019). Ethical issues in big data analytics: A stakeholder perspective. *Communications of the Association for Information Systems, 44*, Article 34.

Speckmann, F. (2021). Web scraping. *Zeitschrift für Psychologie, 229*(4), 241—244.

The Economist. (2017, May 6). The world's most valuable resource is no longer oil, but data. *The Economist*. Available at https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics, 165*(C), 234—246.

Watson, H. J. (2014). Tutorial: Big data analytics: Concepts, technologies, and applications. *Communications of the Association for Information Systems, 34*, Article 65.

White, J. M., & Lidskog, R. (2021). Ignorance and the regulation of artificial intelligence. *Journal of Risk Research, 25*(4), 488—500.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software, 59*(10), 1—23.

Wigan, M. R., & Clarke, R. (2013). Big data's big unintended consequences. *IEEE Computer, 46*(6), 46—53.