The Legality and Ethics of Web Scraping in Archaeology

Jonathan Paige

ABSTRACT

Web scraping, the practice of automating the collection of data from websites, is a key part of how the internet functions, and it is an increasingly important part of the research tool kit for scientists, cultural resources professionals, and journalists. There are few resources intended to train archaeologists in how to develop web scrapers. Perhaps more importantly, there are also few resources that outline the normative, ethical, and legal frameworks within which scraping of archaeological data is situated. This article is intended to introduce archaeologists to web scraping as a research method, as well as to outline the norms concerning scraping that have evolved since the 1990s, and the current state of US legal frameworks that touch on the practice. These norms and legal frameworks continue to evolve, representing an opportunity for archaeologists to become more involved in how scraping is practiced and how it should be regulated in the future.

Keywords: computational archaeology, heritage management, law, web scraping

Web scraping, la práctica de automatizar la recopilación de datos de sitios web, es una parte clave del funcionamiento de Internet y, cada vez más, es una parte importante del conjunto de herramientas de investigación para científicos, profesionales de recursos culturales y periodistas. Hay pocos recursos destinados a capacitar a los arqueólogos sobre cómo desarrollar web scrapers. Quizás lo más importante es que también hay pocos recursos que describan los marcos normativos, éticos y legales dentro de los cuales se sitúa el raspado de datos arqueológicos. Este documento tiene como objetivo presentar a los arqueólogos el web scraping como método de investigación, así como delinear las normas relacionadas con el scraping que han evolucionado desde la década de 1990 y el estado actual de los marcos legales de los Estados Unidos que tocan esta práctica. Estas normas y marcos legales continúan evolucionando, lo que representa una oportunidad para que los arqueólogos se involucren más en cómo se practica y regula el raspado en el futuro.

Palabras clave: arqueología computacional, gestión del patrimonio, ley, extracción de datos web

Web scraping is a key part of how the internet functions, and it is an increasingly important part of the tool kit for scientists, cultural resources professionals, and journalists interested in collecting quantities of data impractical to collect by hand (Dogucu and Çetinkaya-Rundel 2021; Luscombe et al. 2022). Furthermore, as big data and macroarchaeological research become more fully developed (Perreault 2019), and as open science initiatives and online repositories of archaeological data become more established, web scraping and similar computational approaches could play an important—and even necessary—role in the aggregation of large comparative datasets. However, guidance about web scraping is lacking in the field of archaeology. Web scraping is regulated mainly by norms that developed from the mid-1990s through the present day, and through public policy and the legal system (Gold and Latonero 2018; Krotov et al. 2020; Sobel 2021). Both sets of frameworks have seen substantial change, and the legal framework surrounding scraping is hazy. Furthermore, archaeologists are bound by another set of ethical guidelines tailored to how we collect, analyze, and report archaeological data. There is little published work on how scraping relates to these legal and ethical frameworks within archaeology. As a result, guidance about whether, where, when, and how one should scrape is ambiguous or absent.

This article outlines what scraping is, how to do it, and the legal and ethical frameworks within which scraping is situated in the field of archaeology in the United States. As this discussion will highlight, the legal and ethical status of the practice by archaeologists, and by social scientists more generally, is often more ambiguous than would be the case for other kinds of data collection practices. However, that ambiguity is a good opportunity for researchers to become more involved in shaping how this practice is performed and regulated (Luscombe et al. 2022). This article is intended as a small step toward helping archaeologists legally and ethically incorporate this powerful method into their research tool kits as well as providing a framework for archaeologists to critically evaluate work that involves web scraping.

Advances in Archaeological Practice 12(2), 2024, pp. 98–106

DOI:10.1017/aap.2023.42

The open science and open data philosophies that began to mature in the early 2000s influenced the development of web scraping for scientific research. This period saw efforts to organize alternatives to publishing systems funded by institutional subscriptions and to make research products and data publicly available. The products of these efforts include the development of public copyright licenses—such as the Creative Commons License, the Public Library of Open Science, and other open-access research publications—formalized definitions of "open data" and "open access," and public digital repositories for research products including datasets (Harnad 2005; Laakso et al. 2011). Those efforts, in part, resulted in changes to federal policies to encourage data sharing (Sheehan 2015). In the field of archaeology, repositories such as the Digital Archaeological Record (tDAR) developed to fill a need for a public repository for publicly funded archaeological research (McManamon et al. 2017). As these repositories and similar products of archaeological research become more deeply entrenched in the practice of archaeology, and as an appetite for large-scale comparative and synthetic work based on archaeological data increases (Ortman and Altschul 2023; Perreault 2019), the potential for web scrapers and similar kinds of computational methods as important research tools significantly expands (McManamon et al. 2017; Ortman and Altschul 2023).

WHAT SCRAPING IS AND HOW TO DO IT

Archaeologists and other researchers interested in studying how the human past is viewed, studied, and commercialized increasingly rely on web scraping and similar computational tools as the method of gathering data (Daems 2020; Graham et al. 2020; Hashemi and Waddell 2022; Kintigh 2015; Marwick 2014; Richardson 2019; Wilson et al. 2022). Many of these approaches focus on studying how issues relating to archaeological practice are discussed on social media platforms (Marwick 2014; Richardson 2019), while others focus on investigating the online trade of illicit antiquities (Hashemi and Waddell 2022). However, as more and more archaeological data are hosted on websites, web scraping holds greater and greater potential as a powerful data collection tool. Still, despite the power of this method, archaeologists rarely incorporate web scraping into their research.

Scraping involves the automated collection of data by a computer program from websites where that data was intended to be read or collected by humans. People read PDFs, copy and paste elements of spreadsheets, or browse images: all forms of data presentation that are tailored for human consumption. These kinds of formats are unlike the data structures used to exchange information between computers, which are tightly structured, as unambiguous as possible, and read as nonsense to most people (Wiley 2021). Website data transmission happens through an Application Programming Interface (API; Jünger 2021). APIs represent sets of guidelines that structure how data are requested by one party, and how that request is fulfilled by the second party. Access to a website's API is the ideal means of collecting data from that website because it provides more computationally direct access to data. For example, the online Digital Index of North American Archaeology (DINAA) has guides for requesting data through API calls, making web scraping unnecessary (Wells et al. 2014).

However, not all websites have public APIs that researchers can use. Scraping is a fallback option in such cases.

Although diverse kinds of computer languages can be used to create websites, all website pages are transmitted in HyperText Markup Language (HTML) from a server to a client's browser (Mowery and Simcoe 2002). HTML is used to build the structure and content of a web page. The HTML representation of this website structure is called a "parse tree," which is made up of hierarchically organized "elements," each of which is bracketed by a tag. For example, " " would contain a paragraph of text. The information contained within a tag is the "content" of that element. Elements often have attributes associated with them to further specify exactly what element they refer to, such as "id='paragraph-2'" or "name = img3." The simplest methods of web scraping involve navigating this parse tree, identifying strings or tags associated with the data we want, and extracting those associated data—be they images, strings, tables, a single number, or other structures. Using regular expressions to identify tags and elements of interest and then copying material from that element is one popular strategy.

Archaeologists interested in web scraping and who already have programming experience are most likely to have skills in R or Python (Marwick et al. 2017). Both programming languages have useful packages that support web scraping. Organizations such as The Programming Historian and Data Carpentry offer workshops and online tutorials that are designed to train researchers with no prior programming experience. In Python, the Beautiful Soup package is commonly used for scraping (Richardson 2023). The package was developed in 2004, and since then, there have been many tutorials published online, and there is an active community of users posting coding issues and solutions. R is the most widely used programming language in the field of archaeology, although it is less commonly used for web scraping. Recently, however, there has been an expansion of packages designed for web scraping—such as rvest—which was released in 2014 (Wickham 2023). Both rvest and Beautiful Soup contain similar functions that can automate the navigation of an HTML parse tree.

The parse tree, in many instances, may not have the data we want to scrape. Clicking on a website link to access a data table, for example, may not result in navigation to another HTML page, with its parse tree containing the table itself. Instead, websites often incorporate calls to objects stored elsewhere to be displayed within the web browser without changing the parse tree. In such situations, we need other kinds of packages with which either rvest or Beautiful Soup can interact. One solution is to navigate the website itself as a user would—that is, clicking links, typing search queries, and copying data through commands submitted to a web browser. Selenium in Python (Muthukadan 2018) and Rselenium in R (Harrison and Kim 2022) are good packages for automating the navigation of websites from an R or Python session, and they allow us to scrape data that are not represented within an HTML parse tree. Such an approach is likely required for most websites with modern interfaces.

SCRAPING NORMS

Because scraping is an automated process, and because of the generally open nature of the internet, an incredible amount of data can be extracted from websites very quickly. Web scrapers

https://doi.org/10.1017/aap.2023.42 Published online by Cambridge University Press

form the foundation of how search engines, such as Google, develop a map of the internet. However, there are many kinds of ways in which web scraping can be misused. Some web scrapers are designed to copy entire websites and rehost pages under new domains for ad revenue. Some web scrapers also can perform the same function as a distributed denial of service (DDoS) attack by automating repeated and very rapid requests on a website, overwhelming its servers. Other scraping techniques may be designed to extract sensitive information about individuals, which then may be sold to a third party (Krotov et al. 2020). Any researcher who can make a web scraper can also make a web scraper that does similar kinds of intrusive damage. Because scrapers can be used for good or ill, they have become the target of regulation not only through legal systems and policymaking but through the development of norms within communities of web designers and scrapers.

One long-standing norm in website design is the Robots Exclusion Protocol, which was developed in the mid-1990s as a means for website owners to communicate to web scraping programs which pages could be and which should not be scraped, as well as information about which web scraping programs are or are not welcome to scrape those pages (Elmer 2008). A website's Robots Exclusion Protocol is provided as a stand-alone page on a website under the address "/robots.txt." The information in a robots.txt may or may not be wholly consistent with a site's terms of service. For example, a site could have no information about web scraping in its terms of service, but the robots.txt could have instructions that are intended to prohibit most forms of web scraping. Adherence to the exclusion protocol is voluntary, and many scrapers do ignore them, although these exclusion standards and the violation or adherence to them have been cited in legal cases involving web scraping (John F. Tamburo, et al., Plaintiffs, v. Steven Dworkin, et al., Defendants. No. 04 C 3317 [N.D. Ill. Nov. 17, 2010]).

Robots.txt files are highly variable, and they have become generally more complex over time given that more and more programs and services have developed to gather data from websites. For example, the robots.txt file associated with eBay in 1998, the early days of that website, included only four lines prohibiting scraping originating from one source (Figures 1 and 2).

As of summer 2023, eBay has a far stricter robot.txt. This page prohibits scraping except by search engines—or scrapers that help generate advertising revenue (Figure 3). Additionally, it supplies a description of the site's philosophy on web scrapers that is written in prose within the robots.txt document. Nonetheless, eBay, in particular, is one of the sites that has seen the most research on cultural heritage and antiquities markets through web scraping, and there are publicly available tools specifically designed to scrape the online auction house. This has helped us to gain compelling insights about what kinds of antiquities are most popular on the site and, as a result, to better understand the role the antiquities market plays in looting and the preservation of the archaeological record (Altaweel 2019; Altaweel and Hadjitofi 2020).

Other norms have developed over the course of the history of the internet that help determine the design and use of web scrapers. For example, attempting to obfuscate one's identity or misrepresenting the reasons why the data are being scraped—especially if doing so is necessary to obtain access to the data—is ethically dubious. Instead, researchers should consider transparent

```
User-agent: *           # directed to all spiders

User-agent: Roverbot    # directed to roverbot.com
Disallow: /
```

FIGURE 1. Robots.txt file associated with eBay.com in the mid-1990s. The only restriction on web scrapers was the exclusion of the service "roverbot."

practices. This is also an opportunity to explain to the domain owner that the scraping process was designed to avoid harm. Journalists, for example, have developed different norms and practices around how to scrape data from websites (Wiley 2021). One common practice to ensure transparency is to include identifying information within the scraper code—mainly a user agent string that outlines the identity of who is enacting the scraping—some information about the reason behind the scraping, contact information, and steps taken to ensure that the scraper does no damage to the functioning of a website. Other steps include throttling the frequency of requests sent to a website (Densmore 2017; Wiley 2021). These minimally invasive and transparent practices are formalized in the R package "polite," which has built-in functions for scanning the robots.txt file of a website to identify if scraping is allowed, for requesting permission to scrape, and for throttling requests to one every few seconds (Perepolkin 2023).

LEGAL FRAMEWORKS

Researchers, before performing a data-scraping project, should familiarize themselves with federal, state, and local regulations that relate to web scraping, copyright law, and digital trespassing. Currently, there are no laws that are tailored specifically to web scraping. Instead, a patchwork of relevant laws, regulations, and decisions are cited in cases involving web scraping. Until reforms occur, the legal landscape will remain murky (Christensen 2020; Landers et al. 2016; Sellars 2018; Sobel 2021). One caveat for the below discussion: I am not a lawyer. For a fuller discussion of the laws and regulations surrounding web scraping, many of the references in this section are a good start.

Site terms and conditions are one method of proposing the "gatekeeper rights" of the owners of a website (Kadri 2020). Those terms and conditions may explicitly note that automated data collection is not allowed, even if their website makes that data publicly available (Wickham et al. 2023). Terms of use violations, such as scraping publicly available information, could be enough for a company to send a cease-and-desist letter (Kadri 2020). In such cases, there is little evidence for successful criminal cases brought against groups who scraped publicly available information, even though it was against the terms of use of that website. However, more precautions should be taken if access to the data to be scraped requires setting up an account and actively agreeing to a website's terms of service that explicitly bans web scraping (Landers et al. 2016). Any information that is behind a log-in screen or that requires an account and agreement to a terms of service to access is more likely to have stronger protections under federal laws (Landers et al. 2016; Macapinlac 2019).

https://doi.org/10.1017/aap.2023.42 Published online by Cambridge University Press

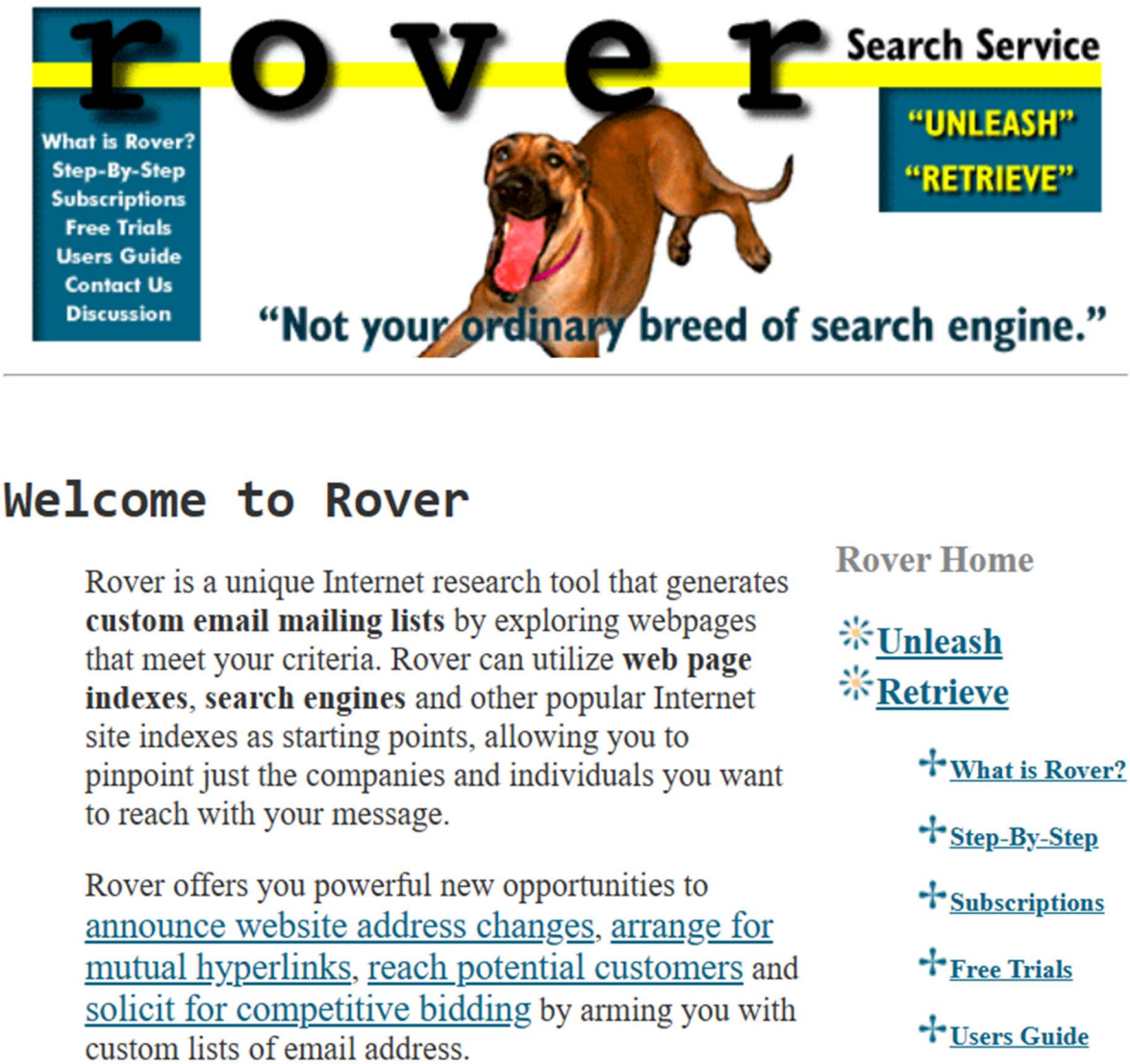

FIGURE 2. RoverBot.com website as it stood in December 1996. This was the only web-scraping program that eBay.com disallowed from scraping data on its website in 1998.

In the United States, the main law that gives shape to government policy regarding web scraping is the Computer Fraud and Abuse Act (CFAA; Christensen 2020; Sellars 2018). The CFAA was passed in 1986, well before public use of the internet was commonplace. It outlines the legal framework surrounding how computers are accessed and used. Mainly, it was intended to protect commercial property rights, and sensitive data hosted on the computers of government agencies. For example, if a disgruntled employee released a computer virus on a company's network, destroying financially valuable data, that would be a crime under the CFAA. In contrast, if that released virus did no economic damage, then whether or not that action would constitute a crime is less clear under the CFAA (Roach and Michiels 2006). The CFAA also vaguely delineates web-scraping practices that could be illegal. For example, if private data not intended for the public were somehow scraped from a website, that could constitute "unauthorized access" and fall under the CFAA (Krotov et al. 2020; Sobel 2021). The notoriously vague wording of the CFAA also means that acts such as lying about one's age on the internet could constitute a federal crime, and even scraping of publicly accessible data could be construed as a federal crime (Macapinlac 2019).

The uncertainty around how scraping relates to legal frameworks leads to a lack of predictability about what kinds of actions will be charged as federal crimes. This uncertainty has led to calls to reform the CFAA and other laws. One of the major historical events that illustrates this was the federal legal action brought against Aaron Swartz in 2011. Swartz was a computer scientist who developed the RSS feed, Markdown, and the Creative Commons license. He was also a leading advocate for the open availability of scientific data and research on the internet. Swartz was indicted under the Computer Fraud and Abuse Act under allegations that he used the free access of MIT's institutional JSTOR subscription

https://doi.org/10.1017/aap.2023.42 Published online by Cambridge University Press

```
## BEGIN FILE ###
#
# allow-all
# DR
#
# The use of robots or other automated means to access the eBay site
# without the express permission of eBay is strictly prohibited.
# Notwithstanding the foregoing, eBay may permit automated access to
# access certain eBay pages but soley for the limited purpose of
# including content in publicly available search engines. Any other
# use of robots or failure to obey the robots exclusion standards set
# forth at <https://www.robotstxt.org/orig.html> is strictly
# prohibited.
#
# v18.5_COM_May_2023
### DIRECTIVES ###

User-agent: *
Disallow: /*_kw
Disallow: /*?maspect
Disallow: /*modules=SEARCH_REFINEMENTS_MODEL_V2
Disallow: /*redirect=mobile
Disallow: /*redirect%3Dmobile
Disallow: /*rt%3Dnc
Disallow: /*rt=nc
Disallow: /*src=urllib
Disallow: /?SSOWebDispatcher=&tg=web&ru={{ru}}
```

FIGURE 3. The first fraction of 475 lines of the robots.txt file associated with eBay.com as of July 2023. Most kinds of web scraping are disallowed.

on the public MIT network to download scientific papers en masse using a scraping program. Although neither MIT nor JSTOR pushed for his prosecution, and although Swartz had not then shared those files with the public, federal prosecutors brought charges that included wire fraud and computer fraud. Swartz took his own life in 2013. There was significant bipartisan backlash against the Justice Department's handling of the case, and the case also further galvanized calls for open access to scientific data both broadly and within the field of archaeology (Kansa et al. 2013). Although reforms to the CFAA were drafted after the Swartz case, none were passed into law. Nonetheless, this case likely had an impact on future applications of the CFAA. Interpretation of the law has continued to evolve, and subsequently, there have been a few other instances of federal charges being brought against researchers and individuals who built web scrapers to collect publicly available data for research or scientific purposes. Instead, CFAA cases involving web scraping tend to revolve around business disputes (Macapinlac 2019).

The reach of the CFAA has become slightly shorter as legal cases continue to be decided (Christensen 2020). For example, one of the recent and higher-profile web-scraping cases brought before the Ninth Circuit Court of Appeals was hiQ Labs Incorporated v. LinkedIn Corporation (938 F.3d 985 [9th Cir. 2019]). The Ninth Circuit evaluated whether LinkedIn could cite the CFAA in a case against the company hiQ, which had been scraping information LinkedIn users had placed on their public profiles. In the end, the Ninth Circuit held that scraping publicly available information does not violate the CFAA (938 F.3d 985 [9th Cir. 2019]; Christensen 2020; Sobel 2021). That, however, does not preclude similar cases being brought under other regulations that focus on intellectual property or trespass, for example (938 F.3d 985 [9th Cir. 2019]; Sobel 2021).

The Digital Millennium Copyright Act of 1998 is one example of a federal copyright law intended to afford companies with copyrighted digital work protections against others republishing or repurposing their works, especially if that reuse is for profit (Lawrence and Ehle 2019). Reuse and reproduction of online materials for the sake of research is more likely to fall under "fair use" exclusions to copyright law (Myers 2022). However, if scraping involves gathering massive amounts of data and rehosting that data in some way with minimal modification, or if scraping is performed in such a way that it has a negative economic impact on a website or company, this could increase the likelihood of successful legal action (Lawrence and Ehle 2019; Liu and Davis 2015).

Common law, or tort law, also can provide a basis for civil cases that could be brought against people who implement web scrapers. A tort refers to some action that causes a claimant some loss or harm. Trespass to chattels is one example of such a tort in civil law (Sobel 2021). Historically, trespass to chattels is a portion of tort law that serves as a basis to bring civil action against individuals who interfere with another's possessions—or "chattel"—through taking those possessions, inhibiting access to them, or

https://doi.org/10.1017/aap.2023.42 Published online by Cambridge University Press

destroying them, for example (Quilter 2002). Trespass to chattels is often used to bring civil cases against sources of spam on the internet. One of the first such cases—CompuServe Inc. v. Cyber Promotions Inc.—involved CompuServe arguing that the bulk digital contact that originated from Cyber Promotions was sufficiently damaging to constitute a trespass to chattel (Graham 1997; Quilter 2002). The courts ruled in favor of CompuServe and opened the door for trespass to chattel cases to be brought against others, even if there were only indirect costs or damages that resulted from the trespass (Quilter 2002). Trespass to chattel is cited in several web-scraping cases as well (O'Reilley 2007), and courts generally appeared willing to rule in favor of companies that bring trespass to chattel charges, even without strong evidence of economic damages caused by scraping (Quilter 2002). This provides another viable avenue for website owners to restrict access to publicly available data that otherwise are more difficult to restrict through either copyright law, such as the DMCA, or digital trespass and fraud laws, such as the CFAA (O'Reilly 2007; Quilter 2002).

In summary, in the case of most research projects that involve scraping of publicly available information, there is a low risk of criminal liability but some risk of civil liability. Given the legal ambiguity surrounding the practice, one strategy is to avoid scraping altogether. However, being uninvolved in scraping as a field also leaves archaeologists and heritage professionals in a place where they cannot influence how the practice is employed and regulated in the future. That same ambiguity has also not stopped web scraping from becoming widespread in business, journalism, and other scientific fields (Baranetsky 2018; Kirkpatrick 2015; Wiley 2021). This is largely because the method can provide economic, scientific, and public benefits that arguably outweigh risks that stem from the ambiguity in the legal framework.

ARCHAEOLOGICAL ETHICS

Over the course of the twentieth century in the United States, there was a transition from the archaeological record being unprotected and unmanaged by public entities to the modern condition where archaeological sites are protected and managed to abide by not only legal obligations but professional norms and ethics focused on site preservation (King and Lyneis 1978; Society for American Archaeology 2016). Since the passage of the Antiquities Act in 1906, the federal government has taken an explicit role in the management of archaeological resources on public lands (Colwell-Chanthaphonh 2005; King and Lyneis 1978). The Archaeological Resources Protection Act of 1979 further outlined the role the government must play in protecting sites or subjecting them to minimal damage during analysis (Northey 1982). Among those new protections were provisions to prevent site location data from becoming accessed by the broader public. This kind of sensitive locational information can be easily collected en masse if an archaeologist has access to websites that store it. As digital methods advance, it is important to continually revisit how our practices relate to our ethical and legal obligations (Dennis 2020; Richardson 2018). These kinds of legal, ethical, and normative obligations should be kept in mind during any attempt to scrape large amounts of data about the archaeological record.

Precautions must be taken when gathering either locational information using a web scraper or any other information that could make it much easier to locate—and damage—archaeological sites. One strategy is to engage in some form of obfuscation of the true site locations (Anderson and Horak 1995; Robinson et al. 2019; Smith 2020). A popular strategy is to summarize the locations of sites analytically based on which county they fall within. Another viable strategy is to resample a new site location from within a certain radius of the reported site location (Smith 2020). These steps are best performed at the same time as the scraping to ensure that the obfuscated location, rather than the true site location, is stored at any point. This is to prevent any researcher from having thousands of precise site locations on a personal or work computer. Even if those raw data are never meant to be shared widely, it is not good practice to retain precise locational data for no good reason, given that it could be leaked, unintentionally shared, or hacked. Strategies such as saving only obfuscated site coordinates throughout scraping help to mitigate that risk.

Furthermore, archaeologists interested in public perceptions of archaeology—or other questions that involve gathering data from or about living people—must also ensure that the rights and welfare of those people are protected. Web scraping, as outlined above, can allow one individual to collect massive amounts of data about individuals from public websites, much of which may result in individuals being indirectly identified even if names and other sensitive forms of personally identifiable information are not collected. In 2013, the Department of Health and Human Services updated its recommendations and guidance specifically for internet-based research, including data gathered through web scraping (Secretary's Advisory Committee on Human Research Protections 2013). This updated guidance outlined a framework relating potentially sensitive information hosted on websites to the "basic ethical principles" of human subject research outlined in the *Belmont Report: Respect for Persons, Beneficence, and Justice* (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1979). The boundary between public information (data that individuals should not expect to be kept private) and private information (such as medical, educational, and financial records) is hazy on the internet. Many users, for example, may not be fully aware that the information they provide on a public website is likely to be observed and recorded for scientific research, even if that data does ostensibly qualify as "public." The beneficence principle in the *Belmont Report* serves to temper a broad treatment of all publicly available information on the internet as ethically "public" given that many users may be operating under an assumption that their data will not be widely spread. In some online communities, there may be a stronger shared expectation of privacy, and the Advisory Committee's recommendation in such an instance is to be aware of and respect those expectations (Secretary's Advisory Committee on Human Research Protections 2013). The 2013 guidance also discusses concerns about the kinds of research studies that qualify as interventions, the kinds of observations that qualify as observations of public behavior, and the argued characteristics of sites that should be considered analogous to public places. A full discussion of these ideas is beyond the scope of this article, but they should be carefully considered when scraping intersects with human subjects research. In the United States, any scraping work that involves studying human subjects must be discussed with Institutional Review Boards (IRBs), regardless of the investigators' perception of risk to those human subjects. As is the case with sensitive site locational information, sensitive

https://doi.org/10.1017/aap.2023.42 Published online by Cambridge University Press

information—including personally identifiable information—should not be stored unless necessary. When it must be, such data should be stored securely, encrypted in a secure server, or both.

DISCUSSION AND CONCLUSIONS

Web scraping is a useful method of sampling from an ever-deepening pool of data hosted on the internet. However, there are many factors to consider when deciding whether and how to build and implement a web scraper. In some cases, we may be more likely to argue that scraping publicly available information from a website is in the public interest and of scientific value, even if the site is explicit in requesting that no web scraping be performed at all (Luscombe et al. 2022). We might, for example, be interested in systematically assessing how discussions about the archaeological record and prehistory have changed over the past few decades by performing a textual analysis of posts on White-supremacist message boards. Although this may be prohibited in the site terms of use or in the robots.txt associated with the message board, such a study may have scientific value and could be in the public interest. It would help us better understand the long-identified relationship between archaeological findings and White nationalism (Hakenbeck 2019). So should researchers always adhere to company requests? Researchers should consider both their justifications for proceeding—whether the data are public, whether the users of the website may expect privacy, and whether scraping will hurt those websites—and input from IRBs.

In contrast, many websites that contain archaeological data do not have information in their terms of use or a robots.txt to give guidance about expectations for the use of web scrapers. For example, the Texas Historical Commission site atlas does not, nor does it have any guidance in its terms of use that directly relates to the use of web scrapers. The lack of guidance from websites should not be considered a license to scrape in whatever way one wishes. In cases like this, researchers should still—at a minimum—identify themselves, focus on targeted collection, and find ways to obfuscate more sensitive information, especially site locational data. Heritage professionals who are building digital repositories and digital interfaces that provide large amounts of digital data—including State Historic Preservation Offices that provide database access to professional archaeologists—should keep in mind the use of web scrapers in the design of those websites. Appropriate robots.txt and discussion in terms of service surrounding the use of web scrapers should help provide an additional layer of protection for more sensitive data.

In summary, there are three broad ways to look at the practice of web scraping. The first is to, out of an abundance of caution, avoid the practice. This circumvents any potential legal or ethical issues outlined above. Another is a no-holds-barred approach, where entire datasets are extracted through whatever means necessary and hosted with no modification, including sensitive data. Researchers should take into account all the issues raised above, norms surrounding scraper design, archaeological ethics, the legal system, the stated desires of website owners, and the expectations of the users of those sites. Researchers should use their best judgment while being aware of the current and changing legal and ethical climates within which this kind of work is situated (Landers et al. 2016; Luscombe et al. 2022). By engaging in this practice, archaeologists and other cultural heritage professionals can then become a part of the community that makes decisions about how scraping should be performed and how it should be regulated in the future.

Acknowledgments

I thank John Potter for helpful comments on the legal aspects of this article, and two anonymous reviewers, whose comments helped strengthen it. This project required no permit and involved no archaeological data collection.

Funding Statement

This project was not funded through a grant.

Data Availability Statement

No data were collected or reported in this work.

Competing Interests

The author declares none.

REFERENCES CITED

Altaweel, Mark. 2019. Data for eBayScraper NLP Tool. *UCL Discovery*. https://discovery.ucl.ac.uk/id/eprint/10079023/, accessed March 12, 2023.

Altaweel, Mark, and Tasoula Georgiou Hadjitofi. 2020. The Sale of Heritage on eBay: Market Trends and Cultural Value. *Big Data & Society* 7(2). https://doi.org/10.1177/2053951720968865.

Anderson, David G., and Virginia Horak. 1995. *Archaeological Site File Management: A Southeastern Perspective*. Interagency Archeological Services Division, National Park Service, Southeast Regional Office, Atlanta, Georgia.

Baranetsky, D. Victoria. 2018. Data Journalism and the Law. *Columbia Journalism Review*, September 19. https://www.cjr.org/tow_center_reports/data-journalism-and-the-law.php/, accessed July 30, 2023.

Christensen, Jennie E. 2020. The Demise of the CFAA in Data Scraping Cases. *Notre Dame Journal of Law, Ethics & Public Policy* 34:529–547.

Colwell-Chanthaphonh, Chip. 2005. The Incorporation of the Native American Past: Cultural Extermination, Archaeological Protection, and the Antiquities Act of 1906. *International Journal of Cultural Property* 12(3):375–391. https://doi.org/10.1017/S0940739105050198.

Daems, Dries. 2020. A Review and Roadmap of Online Learning Platforms and Tutorials in Digital Archaeology. *Advances in Archaeological Practice* 8(1):87–92. https://doi.org/10.1017/aap.2019.47.

Dennis, L. Meghan. 2020. Digital Archaeological Ethics: Successes and Failures in Disciplinary Attention. *Journal of Computer Applications in Archaeology* 3(1):210–218. https://doi.org/10.5334/jcaa.24.

Densmore, James. 2017. Ethics in Web Scraping. Electronic document, https://towardsdatascience.com/ethics-in-web-scraping-b96b18136f01, accessed July 31, 2023.

Dogucu, Mine, and Mine Çetinkaya-Rundel. 2021. Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities. *Journal of Statistics and Data Science Education* 29(sup1):S112–S122. https://doi.org/10.1080/10691898.2020.1787116.

Elmer, Greg. 2008. Exclusionary Rules? The Politics of Protocols. In *Routledge Handbook of Internet Politics*, edited by Andrew Chadwick and Philip Howard, pp. 376–383. Routledge, London.

Gold, Zachary, and Mark Latonero. 2018. Robots Welcome? Ethical and Legal Considerations for Web Crawling and Scraping. *Washington Journal of Law, Technology & Arts* 13(3):275–312.

Graham, James. 1997. CompuServe Inc. v. Cyber Promotions, Inc., 962 F. Supp. 1015 (S.D. Ohio 1997).

https://doi.org/10.1017/aap.2023.42 Published online by Cambridge University Press

Graham, Shawn, Damien Huffer, and Jeff Blackadar. 2020. Towards a Digital Sensorial Archaeology as an Experiment in Distant Viewing of the Trade in Human Remains on Instagram. *Heritage* 3(2):208–227. http://doi.org/10.3390/heritage3020013.

Hakenbeck, Susanne E. 2019. Genetics, Archaeology and the Far Right: An Unholy Trinity. *World Archaeology* 51(4):517–527. https://doi.org/10.1080/00438243.2019.1617189.

Harnad, Stevan. 2005. The Implementation of the Berlin Declaration on Open Access. *D-Lib Magazine* 11(3). https://eprints.soton.ac.uk/260690/2/03harnad.html, accessed September 25, 2023.

Harrison, John, and Ju Yeong Kim. 2022. RSelenium: R Bindings for "Selenium WebDriver." R package version 1.7.7. https://cran.r-project.org/web/packages/RSelenium/, accessed March 20, 2024.

Hashemi, Layla, and Abi Waddell. 2022. Investigating the Online Trade of Illicit Antiquities. In *Antiquities Smuggling in the Real and Virtual World*, edited by Layla Hashemi and Louise Shelley, pp. 218–239. Routledge, New York.

Jünger, Jakob. 2021. A Brief History of APIs. In *Handbook of Computational Social Science*, Vol. 2, edited by Uwe Engel, Anabel Quan-Haase, Sunny Liu, and Lars Lyberg, pp. 17–32. Routledge, London.

Kadri, Thomas. 2020. Digital Gatekeepers. *Texas Law Review* 99(5):951–1003. https://doi.org/10.2139/ssrn.3665040.

Kansa, Eric C., Sarah Whitcher Kansa, and Lynne Goldstein. 2013. On Ethics, Sustainability, and Open Access in Archaeology. *SAA Archaeological Record* 13(4):15–22.

King, Thomas F., and Margaret M. Lyneis. 1978. Preservation: A Developing Focus of American Archaeology. *American Anthropologist* 80(4):873–893. https://doi.org/10.1525/aa.1978.80.4.02a00060.

Kintigh, Keith W. 2015. Extracting Information from Archaeological Texts. *Open Archaeology* 1(1). https://doi.org/10.1515/opar-2015-0004.

Kirkpatrick, Keith. 2015. Putting the Data Science into Journalism. *Communications of the ACM* 58(5):15–17.

Krotov, Vlad, Leigh Johnson, and Leiser Silva. 2020. Legality and Ethics of Web Scraping. *Communications of the Association for Information Systems* 47:539–563. https://doi.org/10.17705/1CAIS.04724.

Laakso, Mikael, Patrik Welling, Helena Bukvova, Linus Nyman, Bo-Christer Björk, and Turid Hedlund. 2011. The Development of Open Access Journal Publishing from 1993 to 2009. *PLoS ONE* 6(6):e20961. https://doi.org/10.1371/journal.pone.0020961.

Landers, Richard N., Robert C. Brusso, Katelyn J. Cavanaugh, and Andrew B. Collmus. 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.

Lawrence, J. Alexander, and Kristina Ehle. 2019. Combatting Unauthorized Webscraping – The Remaining Options in the United States for Owners of Public Websites despite the Recent hiQ Labs v. LinkedIn Decision. *Computer Law Review International* 20(6):171–174.

Liu, Philip H., and Mark Edward Davis. 2015. Web Scraping – Limits on Free Samples. *Landslide* 8:54–63.

Luscombe, Alex, Kevin Dick, and Kevin Walby. 2022. Algorithmic Thinking in the Public Interest: Navigating Technical, Legal, and Ethical Hurdles to Web Scraping in the Social Sciences. *Quality & Quantity* 56(3):1023–1044. https://doi.org/10.1007/s11135-021-01164-0.

Macapinlac, Tess. 2019. The Legality of Web Scraping: A Proposal. *Federal Communications Law Journal* 71(3):399–422.

Marwick, Ben. 2014. Discovery of Emergent Issues and Controversies in Anthropology Using Text Mining, Topic Modeling, and Social Network Analysis of Microblog Content. In *Data Mining Applications with R*, edited by Yanchang Zhao and Yonghua Cen, pp. 63–93. Elsevier, Waltham, Massachusetts.

Marwick, Ben, Jade d'Alpoim Guedes, C. Michael Barton, Lynsey Bates, Michael Baxter, Andrew Bevan, Elizabeth Bollwerk, et al. 2017. Open Science in Archaeology. *SAA Archaeological Record* 17(4):8–14.

McManamon, Francis P., Keith W. Kintigh, Leigh Anne Ellison, and Adam Brin. 2017. tDAR: A Cultural Heritage Archive for Twenty-First-Century Public Outreach, Research, and Resource Management. *Advances in Archaeological Practice* 5(3):238–249. https://doi.org/10.1017/aap.2017.18.

Mowery, David C., and Timothy Simcoe. 2002. Is the Internet a US Invention?—An Economic and Technological History of Computer Networking. *Research Policy* 31(8–9):1369–1387. https://doi.org/10.1016/S0048-7333(02)00069-0.

Muthukadan, Baiju. 2018. Selenium with Python — Selenium Python Bindings 2. Electronic document, https://selenium-python.readthedocs.io/, accessed April 12, 2023.

Myers, Gary. 2022 Muddy Waters: Fair Use Implications of Google v. Oracle America, Inc. *Northwest Journal of Technology and Intellectual Property* 19(2):155–190.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Electronic document, https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html, accessed September 9, 2023.

Northey, Lorrie D. 1982. The Archaeological Resources Protection Act of 1979: Protecting Prehistory for the Future. *Harvard Environmental Law Review* 6:61–116.

O'Reilly, Sean. 2007. Nominative Fair Use and Internet Aggregators: Copyright and Trademark Challenges Posed by Bots, Web Crawlers and Screen-Scraping Technologies. *Loyola Consumer Law Review* 19:273–288.

Ortman, Scott G., and Jeffrey H. Altschul. 2023. What North American Archaeology Needs to Take Advantage of the Digital Data Revolution. *Advances in Archaeological Practice* 11(1):90–103. https://doi.org/10.1017/aap.2022.42.

Perepolkin, Dmytro. 2023. polite: Be nice on the web. https://github.com/dmi3kno/polite, accessed July 1, 2023.

Perreault, Charles. 2019. *The Quality of the Archaeological Record*. University of Chicago Press, Chicago.

Quilter, Laura. 2002. The Continuing Expansion of Cyberspace Trespass to Chattels. *Berkeley Technology Law Journal* 17:421–433.

Richardson, Leonard. 2023. Beautiful Soup 4.12.0 documentation. Python package version 4.12.0. https://www.crummy.com/software/BeautifulSoup/bs4/doc/, accessed March 20, 2024.

Richardson, Lorna-Jane. 2018. Ethical Challenges in Digital Public Archaeology. *Journal of Computer Applications in Archaeology* 1(1):64–73.

Richardson, Lorna-Jane. 2019. Using Social Media as a Source for Understanding Public Perceptions of Archaeology: Research Challenges and Methodological Pitfalls. *Journal of Computer Applications in Archaeology* 2(1):151–162. https://doi.org/10.5334/jcaa.39.

Roach, George, and William J. Michiels. 2006. Damages Is the Gatekeeper Issue for Federal Computer Fraud. *Tulane Journal of Technology & Intellectual Property* 8:61–85.

Robinson, Erick, Christopher Nicholson, and Robert L. Kelly. 2019. The Importance of Spatial Data to Open-Access National Archaeological Databases and the Development of Paleodemography Research. *Advances in Archaeological Practice* 7(4):395–408. https://doi.org/10.1017/aap.2019.29.

Secretary's Advisory Committee on Human Research Protections. 2013. Considerations and Recommendations Concerning Internet Research and Human Subjects Research Regulations, with Revisions. Electronic document, https://www.hhs.gov/ohrp/sites/default/files/ohrp/sachrp/mtgings/2013%20March%20Mtg/internet_research.pdf, accessed September 25, 2023.

Sellars, Andrew. 2018. Twenty Years of Web Scraping and the Computer Fraud and Abuse Act. *Boston University Journal of Science & Technology Law* 24:372–415.

Sheehan, Beth. 2015. Comparing Digital Archaeological Repositories: tDAR Versus Open Context. *Behavioral & Social Sciences Librarian* 34(4):173–213. https://doi.org/10.1080/01639269.2015.1096155.

Smith, Cecilia. 2020. Ethics and Best Practices for Mapping Archaeological Sites. *Advances in Archaeological Practice* 8(2):162–173. https://doi.org/10.1017/aap.2020.9.

Sobel, Benjamin L. W. 2021. A New Common Law of Web Scraping. *Lewis & Clark Law Review* 25:147–207.

Society for American Archaeology. 2016. Principles of Archaeological Ethics. Electronic document, https://documents.saa.org/container/docs/default-source/doc-careerpractice/saa_ethics.pdf?sfvrsn=75f1b83b_4, accessed September 5, 2023.

Wells, Joshua J., Eric C. Kansa, Sarah W. Kansa, Stephen J. Yerka, David G. Anderson, Thaddeus G. Bissett, Kelsey Noack Myers, and R. Carl DeMuth. 2014. Web-Based Discovery and Integration of Archaeological Historic Properties Inventory Data: The Digital Index of North American

https://doi.org/10.1017/aap.2023.42 Published online by Cambridge University Press

Archaeology (DINAA). *Literary and Linguistic Computing* 29(3):349–360. https://doi.org/10.1093/llc/fqu028.

Wickham, Hadley. 2023. rvest: Wrappers around the xml2 and httr Packages to Make It Easy to Download, Then Manipulate, HTML and XML. R package version 1.0.4. https://cran.r-project.org/web/packages/rvest/index.html, accessed March 20, 2024.

Wickham, Hadley, Mine Cetinkaya-Rundel, and Garrett Grolemund. 2023. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 2nd ed. O'Reilly, Beijing.

Wiley, Sarah K. 2021. The Grey Area: How Regulations Impact Autonomy in Computational Journalism. *Digital Journalism* 11(6):899–905.

Wilson, Andrew S., Vincent Gaffney, Chris Gaffney, Eugene Ch'ng, Richard Bates, Elgidius B. Ichumbaki, Gareth Sears, et al. 2022. Curious Travellers: Using Web-Scraped and Crowd-Sourced Imagery in Support of Heritage Under Threat. In *Visual Heritage: Digital Approaches in Heritage Science*, edited by Eugene Ch'ng, Henry Chapman, Vincent Gaffney, and Andrew S. Wilson, pp. 51–65. Springer Series on Cultural Computing. Springer International Publishing, Cham, Switzerland.

AUTHOR INFORMATION

Jonathan Paige ■ Department of Anthropology, University of Missouri, Columbia, MO, and Center for Archaeological Research, University of Texas, San Antonio, TX (jonathan.n.paige@gmail.com, corresponding author)

https://doi.org/10.1017/aap.2023.42 Published online by Cambridge University Press