Communications of the Association for Information Systems

Volume 47 Paper

12-10-2020

Tutorial: Legality and Ethics of Web Scraping

Vlad Krotov
Murray State University, vkrotov@murraystate.edu

Leigh Johnson
Murray State University

Leiser Silva
University of Houston

Recommended Citation
Krotov, V., Johnson, L., & Silva, L. (2020). Tutorial: Legality and Ethics of Web Scraping. Communications of the Association for Information Systems, 47, pp-pp. https://doi.org/10.17705/1CAIS.04724

Communications of the Association for Information Systems

Research Article DOI: 10.17705/1CAIS.04724 ISSN: 1529-3181

Tutorial: Legality and Ethics of Web Scraping

Vlad Krotov
Murray State University
vkrotov@murraystate.edu

Leigh Redd Johnson
Murray State University

Leiser Silva
University of Houston

Abstract:

Researchers and practitioners often use various tools and technologies to automatically retrieve data from the Web (often referred to as Web scraping) when conducting their projects. Unfortunately, they often overlook the legality and ethics of using these tools to collect data. Failure to pay due attention to these aspects of Web Scraping can result in serious ethical controversies and lawsuits. Accordingly, we review legal literature together with the literature on ethics and privacy to identify broad areas of concern together with a list of specific questions that researchers and practitioners engaged in Web scraping need to address. Reflecting on these questions and concerns can potentially help researchers and practitioners decrease the likelihood of ethical and legal controversies in their work.

Keywords: Big Data, Web Data, Web Scraping, Web Crawling, Law, Legality, Ethics, Privacy.

This manuscript underwent editorial review. It was received 08/09/2019 and was with the authors for nine months for two revisions. The Associate Editor chose to remain anonymous.

1 Introduction

In the past, researchers and practitioners found it difficult and costly to obtain data for their projects (Munzert, Rubba, Meißner, & Nyhuis, 2015). In contrast, today, increasingly digitized and virtualized social processes have resulted in zettabytes (billions of gigabytes) of available data on the Web (Cisco, 2016). This data provides a granular, real-time representation of numerous processes, relationships, and interactions in the socio-material space (Krotov & Tennyson, 2018). Thus, academic researchers have ample opportunities for answering new and old research questions with more rigor, precision, and timelines (Constantiou & Kallinikos, 2015). Practitioners can leverage this data to better understand their customers, formulate strategies based on their findings, and, ultimately, improve organizational performance (Ives, Palese, Rodriguez, 2016).

Unfortunately, harnessing Web data presents serious technical, legal, and ethical challenges. While tools and technologies that one can use to scrape the Web have proliferated in recent years (Munzert et al., 2015), collecting data from the Web remains a legal and ethical "grey area" (Snell & Menaldo, 2016). While one can apply existing legal frameworks to some extent to Web scraping, researchers have largely ignored Web scraping's ethical issues. In this paper, we review the legal and the general information systems (IS) literatures related to Web data, ethics, and privacy to identify broad areas of concern and specific issues that one needs to address when collecting data from the Web using automated tools. Compliance with these legal and ethical requirements can help industry and academic researchers decrease the likelihood that they will encounter legal problems and ethical controversies in their work and, overall, foster research that relies on Web data.

2 Web Scraping Explained

2.1 Big Web Data

The data available on the Web comprises structured, semi-structured, and unstructured quantitative and qualitative data in the form of Web pages, HTML tables, Web databases, emails, tweets, blog posts, photos, videos, and so on (Watson, 2014). Harnessing data on the Web requires one to address various technical issues related to its volume, variety, velocity, and veracity (IBM, 2018).

First, volume measured in zettabytes (billions of gigabytes) often characterizes data on the Web (Cisco, 2016). Second, the vast data repositories available on the Web come in various formats and rely on various technological and regulatory standards (Basoglu & White, 2015). Third, data on the Web does not remain static; actors generate and modify it with extreme velocity. Fourth, veracity also characterizes data on the Web (IBM, 2018). Due to the open, voluntary, and often anonymous interactions on the Web, Web data's availability and quality remain inherently uncertain. Thus, researchers can never be completely sure whether they can or will be able to access the data they need on the Web and whether it will be valid and reliable enough for research purposes (IBM, 2018).

2.2 Web-scraping Processes, Technologies, and Tools

Given big Web data's volume, variety, and velocity, individuals or even large teams of academic researchers or industry data specialists can find it difficult or impossible to collect and organize it (Krotov & Tennyson, 2018; Krotov & Silva, 2018). As such, researchers often resort to various technologies and tools to automate the data-collection and –organization process to some degree. Researchers often refer to this emerging practice of automatically extracting and organizing data from the Web in order to further analyze it as Web scraping (Krotov & Silva, 2018; Krotov & Tennyson, 2018).

Web scraping comprises three primary and intertwined phases: 1) website analysis, 2) website crawling, and 3) data organization (see Figure 1) (Krotov & Silva, 2018; Krotov & Tennyson, 2018). Each phase requires one to understand several Web technologies and at least one popular programming language, such as R or Python. However, these three phases often require at least some human involvement and, thus, one cannot fully automate them. We discuss each phase further below.

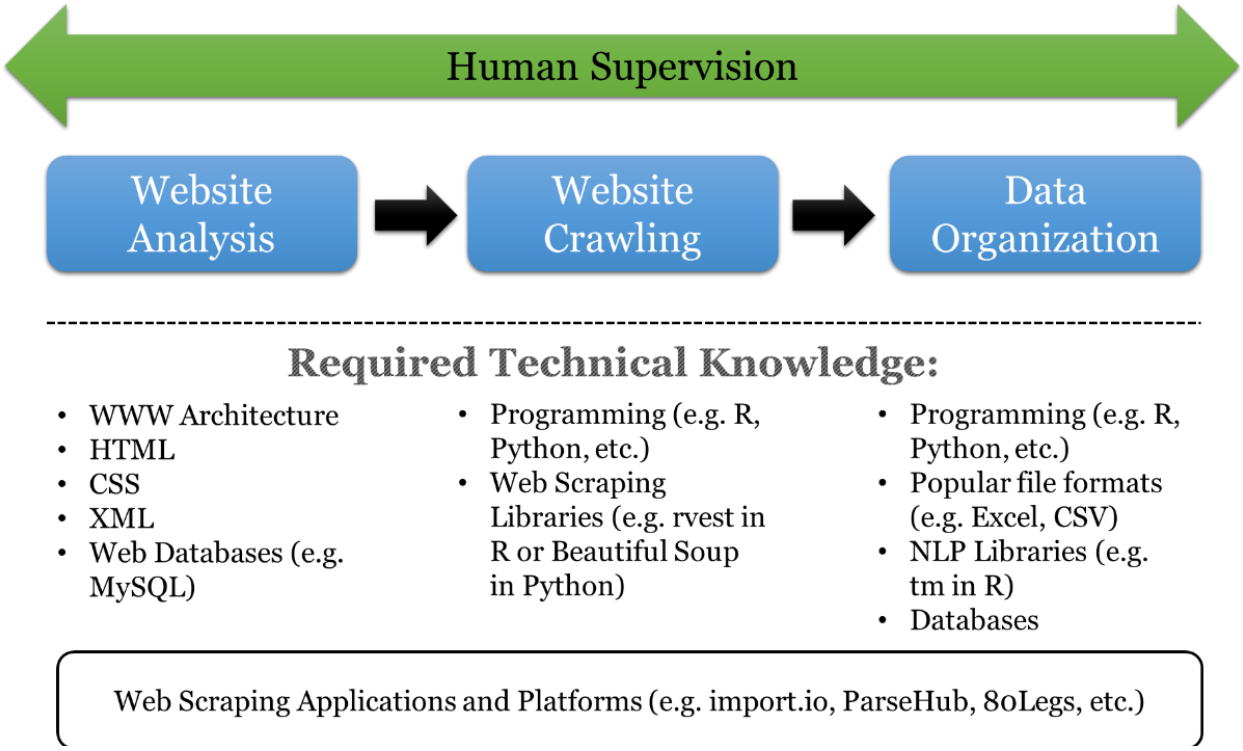


Figure 1. Web Scraping (Adapted from Krotov & Tennyson, 2018; Krotov & Silva, 2018)

Website analysis involves examining a website's, websites', or Web repository's (e.g., an online database) underlying structure to understand how they store data. To do so, one needs to understand the World Wide Web's architecture, mark-up languages (e.g., HTML, CSS, XML, XBRL, etc.), and various Web databases (e.g., MySQL).

Website crawling involves developing and running a script that automatically browses a website and retrieves the needed data. Researchers often develop these crawling applications (or scripts) using such programming languages as R and Python due to their overall popularity in data science and availability in libraries (e.g., "rvest" package in R or Beautiful Soup library in Python) to automatically crawl and parse Web data.

After one parses necessary data from the selected Web repository, one needs to clean, pre-process, and organize it in a way that enables one to further analyze it. Given the volume of data involved, one may also need to adopt a programmatic approach to save time. Many programming languages, such as R and Python, contain natural language processing (NLP) libraries and data-manipulation functions that can help one clean and organize data.

In recent years, "point-and-click" Web-scraping tools that allow one to automate at least some steps in the Web-scraping process without deeply understanding the Web technologies that we discuss above have proliferated. One can use these tools on cloud-based platforms or as standalone desktop applications. We list some such tools in Table 1 (Sagina, 2018).

However, these "point-and-click" Web-scraping tools do not always work as intended. For example, they often miss certain webpages or simply lack the "intelligence" to figure out how to access the needed data points from a website. Given that the Web constitutes an open system with various standards and technologies for delivering Web content and supporting user interactions, we can understand these limitations. Moreover, these standards evolve all the time, and different actors often use and interpret them differently. Overall, the Web contains fluid content: actors generate and modify it at a rapid speed. Thus, even some advanced visual Web-scraping tools unsurprisingly require one to insert custom programming code into them to make them work as intended.

All these technical issues and the fact that many of these tools require hefty subscription fees force many academic and industry researchers to develop their own custom Web-scraping tools using such popular

languages as Python and R. Moreover, recounting how one collected, cleaned, and organized Web data via R or Python computer code in a detailed and unambiguous way can enhance the research protocol that one uses in studies that rely on this Web data (Peng, 2011).

Table 1. Popular Web-scraping Tools (Adapted from Sagina, 2018)

Tool	Description	Notable features
import.io	A visual, cloud-based tool for extracting Web data that individuals and organizations can use	Web-based, user-friendly interface Easy to set up automatic interactions with Web forms and authentication Can schedule data-extraction jobs Can store extracted data in the cloud Can generate data insights via reports, charts, and other data-visualization techniques Can automate data-extraction workflows
Dexi.io	Visual, cloud-based tool positioned as an enterprise solution	Can create robots or agents to extract and clean data Intelligent extraction that makes choices and suggestions to resolve issues related to images, pagination, etc.
Octaparse	Advanced Web-scraping tool with workflow automation based on a Web interface	A mature tool with seven revised versions Intuitive point-and-click interface Supports Web authentication Allows exporting data in multiple formats Supports scheduled crawling
ParseHub	A desktop application with a graphical interface that can access “tricky” Web repositories	Can crawl and scrape typically problematic webpages, such as webpages with nested comments, images, calendars, pop-up windows, AJAX or JavaScript code, etc. Available for multiple platforms, such as Windows, Mac OS X, and Linux
OutwitHub	A desktop tool with an intuitive graphical interface for extracting contacts from the Web.	Can extract links, email addresses, RSS feeds, data tables, or data from databases
FMiner	A Python-based, visual Web-scraping tool for designing Web scraping macros	Can run on Windows and MacOS machines Allows one to embed chunks of Python code to perform descripting actions on a website Supports advanced Web-scraping tasks related to captcha solving and data cleaning
80Legs	Cloud-based tool that overcomes typical deficiencies that Web scrapers experience related to servers limiting automated access to content	Overcomes access rate limiting by rotating among numerous IP addresses Contains pre-built applications for various Web-scraping tasks.

2.3 Addressing Legal and Ethical Issues

In this paper, we argue that, , to produce publishable research or data products, researchers need to not only have knowledge in various tools and technologies for retrieving and organizing data from the Web but also reflect on whether they use Web data in a legal and ethical manner. Court cases that involve disputes over Web data have involved several legal frameworks or broad principles, such as illegal access and use of data, breach of contract, copyright, trespass to chattels, and trade secrets. Researchers should also recognize their actions’ possible ethical consequences in relation to Web data. For example, a finding derived from a research project that relies on data collected from a website may unintentionally compromise individuals’ privacy, violate their rights as research subjects, lead to erroneous decisions, or contribute to bias and discrimination. Web data can also reveal confidential information about organizations that participate in the activities that a website or organization that owns it affords. Furthermore, some ways to use data may reduce a website’s perceived value among its intended audience.

Researchers need to recognize the legal and ethical issues that surround Web scraping. Recognizing potential legal and ethical issues that arise from collecting Web data and taking proactive measures to address these issues can help researchers avoid costly lawsuits or widely publicized ethical controversies

that can damage their reputation. Researchers can waste the time and resources they expend in collecting big data from the Web if they do not collect it in accordance with legal and ethical standards. In this paper, we propose a framework that can guide practitioners and academics on the legality and ethics of collecting Web data, which we call the legality and ethics framework for Web scraping (see Figure 2 later in the paper).

3 Web-scraping Literature

We searched several popular databases for journal papers that contained "Web scraping" or "data scraping" in their title and found that the topic remains a new, emerging phenomenon for two reasons. First, most papers that contained these two phrases appeared only a few years ago. Second, we found fewer than 200 papers that focused explicitly on Web scraping (see Table 2).

Table 2. Web-scraping Literature

Database	Technical papers (n)		Technical total	Non-technical papers (n)		Non-technical total	Total papers
	Tutorials	Solutions		Legal	Ethical		
arXiv	4	3	7	0	0	0	7
EBSCO Academic Search Complete	11	5	16	2	0	2	18
EBSCO Business Source Complete	13	1	14	3	0	3	17
IEEE Xplore	17	86	103	0	0	0	103
JSTOR	0	0	0	2	0	2	2
	Technical papers subtotal (n):		140	Non-technical papers subtotal (n):		7	
	Technical papers (%):		95	Non-technical papers (%):		5	
					Grand total:		**147**

These Web-scraping papers fell into two broad categories: 1) technical, tutorial-like papers on Web scraping and related tools and technologies and 2) non-technical papers that touched on Web scraping's legal and ethical aspects. We found far fewer papers in the second category, which suggests that researchers often overlook social and ethical issues related to Web scraping. We discuss these broad literature categories and representative papers in Sections 3.1 and 3.2. We focus not on rigorously or comprehensively reviewing the Web-scraping literature but on making an analytic argument that we require an integrated, socio-technical approach to Web scraping that considers not only technical but also ethical and legal issues surrounding this emerging practice.

3.1 Technical Papers

The first subcategory of technical papers on Web scraping comprised tutorial-like papers. These papers either discussed technologies and tools that researchers can use for Web Scraping in a particular context or field or outline potential opportunities that Web scraping can offer to researchers in various fields. For example, Krotov and Tennyson (2018) discussed popular mark-up languages that researchers can use for storing financial data on the Web and showed how they can use R language, together with its libraries, to gather, organize, and pre-process financial data from the Web. Neumann, Steinberg, and Schaer (2017) explained how researchers interested in scientometric research can use OXPath technology to harvest metadata from digital library repositories. Beoing and Waddell (2017) showed how researchers can retrieve and analyze data from Craigslist to better understand the rental market in the United States. Kirkpatrik (2015) discussed specific ways in which one can use Web scraping, data mining, and data visualizations in journalism. Collectively, these authors indicated that Web scraping has attracted increasing interest from many scientific fields and industries.

The second subcategory of technical papers related to Web Scraping comprised mostly conference papers that discussed specific technical solutions related to acquiring data from the Web. Computer science researchers authored most of these papers. For example, Hernandez-Suarez at el. (2018) discussed ways in which researchers can bypass Twitter API use restrictions (mostly related to how much

and how often one retrieves data from Twitter). Chaulagain, Pandey, Basnet, and Shakya (2017) proposed an architecture for a cloud-based tool that researchers can use to retrieve big data from the Web. Ujwal, Gaind, Kundu, Holla, and Rungta (2017) proposed a concept of an adaptive Web-scraping tool that adapts to the structural changes of a webpage that contains data that one needs. One can find many other interesting solutions to Web-scraping problems and tools for specific Web-scraping applications in recent conference papers related to Web scraping. Collectively, these papers show that Web scraping constitutes an important and growing area in computer science (e.g., Ujwal et al., 2017).

3.2 Papers Related to Legal and Ethical Issues

We found far fewer papers on Web scraping in the legal and ethical category compared to the technical category (Dreyer & Stockton, 2013; Hirschey, 2014; Snell & Menaldo, 2016; Buchanan, 2017; Krotov & Silva, 2018). Among other things, this literature has concluded that an inherent paradox surrounds Web data, which makes its retrieval and analysis complex from legal and ethical standpoints. On the one hand, the Web's creators intended it to be open and easily accessible to the public. The same openness principle drives many online business models: website owners benefit from a wider user base that can access the data they make available. On the other hand, Web data represents a critical asset for website owners, and, therefore, they need to protect it. Ideally, many website owners would like others to view this data as propriety, which means the individual or entity behind the website that contains this data owns it. However, one cannot easily solidify this ownership since, from a legal standpoint, a website's owner does not necessarily own the data that its users generate (Dreyer & Stockton, 2013).

Either due to all these complexities surrounding Web data ownership or, perhaps, due to the Web-scraping phenomenon's novelty, lawmakers have yet to develop any legislation specific to Web scraping. Instead, legal principles and frameworks that people developed in other eras and different contexts often guide researchers and practitioners. Currently, fundamental legal theories and laws such as "illegal access and use of data", "breach of contract", "copyright infringement", and "trespass to chattels" guide Web scraping (Dreyer & Stockton, 2013; Snell & Menaldo, 2016).

Ethics and the law represent distinct but complimentary notions (Mingers & Walsham, 2010) as legal frameworks usually cover the most obvious ethical issues surrounding a particular practice that one can codify in the form of a relatively unambiguous law (Light & McGrath, 2010). But more complex and subtle ethical controversies often plague Web scraping (Krotov & Silva, 2018). While some legal papers touch on the ethical issues associated with Web scraping, we found no academic paper that has explicitly focused on more subtle ethical issues surrounding Web scraping.

3.3 Addressing the Gaps in the Literature

Based on our literature review, we conclude that researchers often view Web scraping as a technical phenomenon. Most papers that we have discovered tackled technical issues surrounding Web scraping. Few papers touched on the "softer issues" surrounding the practice. While we found quite a few legal papers devoted Web scraping's legality, most had an exploratory nature. They outlined some applicable legal frameworks that [courts] have used in relation to Web scraping and provided some examples of court cases based on these legal frameworks (Gold & Latonero, 2018; Sellars, 2018; Zamora, 2019). We can understand these findings given that Web scraping represents a relatively new, emerging practice and courts have applied various legal theories to Web scraping in an inconsistent manner. Accordingly, individuals engaged in Web scraping cannot easily (or at all) determine their actions' legality. Further, as Web scraping's social acceptability emerges, more subtle legal issues may come into play due to various research projects that rely on Web data. Depending on a study's or a project's goals, one may apply these broad legal frameworks and principles differently.

To address this gap, we thoroughly reviewed the court cases that have involved disputes over Web scraping for the past two decades, analyzed the dispositive facts in each case, and grouped them by similar findings and precedents. In particular, we used the following procedure to identify cases relevant to Web scraping. First, we identified all law journal papers related to Web scraping via a database called HeinOnline. Second, we examined these legal papers to identify the most recent and relevant papers. Subsequently, we identified court cased that these papers discussed. We researched the cases that we identified through "backwards" analysis using Google search engine. We examined both the case history and recent updates to each case using publicly available online sources found via Google. By reviewing and analyzing these court cases, we could not only identify and refine the broad legal issues and

frameworks applicable to Web scraping but also summarize how researchers have applied the various legal theories to Web-scraping practices.

In this review, we also found that the legal literature has often mentioned but rarely explicitly discussed more subtle ethical issues surrounding Web scraping. To the best of our knowledge, we did not find any single paper that explicitly and wholly examined ethical issues surrounding Web scraping. To address this gap, we reviewed IS literature related to Web data collection and ethics. While ethics does not represent a "native" topic in the IS field, we believe that the field has a unique position to adopt a holistic, socio-technical perspective on what we view as a complex, socio-technical phenomenon (i.e., Web scraping). To identify applicable literature, we conducted a backward citation analysis on Mason's (1986) paper, an early and seminal IS paper devoted to ethics in the information age. We scanned the literature to identify several papers that mentioned Web data collection and use. While we did not find many papers, we use them to strengthen and expand our own arguments in relation to ethical issues around Web scraping that originated from our own practical experiences in relation to automated Web data collection.

We used our review of legal and IS literature to identify the most fundamental legal and ethical principles that pertain to Web scraping. We discuss these principles and their arising implications in Section 4.

4 Legality of Web Scraping

Currently, no legislation addresses Web scraping directly; thus, individuals who engage in Web scraping activities do so in an uncertain legal landscape. Currently, related fundamental legal theories and laws, such as "illegal access and use of data", "breach of contract", "copyright infringement", and "trespass to chattels", guide Web scraping (Dreyer & Stockton, 2013; Snell & Menaldo, 2016). We discuss how these fundamental legal theories apply to Web scraping specifically below.

4.1 Illegal Access and Use of Data

Several laws prohibit one from illegally or fraudulently using data that one obtains from Web scraping. The Computer Fraud and Abuse Act (CFAA) and its comparable state laws constitute the predominant legal basis for claims that involve disputes about Web scraping. Among other provisions, the CFAA prohibits the one from intentionally accessing a computer without authorization or in a way that exceeds authorization (18 U.S.C. § 1030(a) (2008)), and the act provides for both civil and criminal penalties (18 U.S.C. § 1030(c) (2008)). Currently, approximately sixty legal opinions have addressed how the CFAA applies to Web scraping. The majority of these opinions address what constitutes unauthorized access under the law. This variety in opinions suggests that courts remain divided on this issue, and they have reached little consensus over the last two decades about how the CFAA applies to Web scraping (Sellars, 2018). Still, recent cases provide some guidance to researchers who employ Web scraping to collect data about their potential civil and criminal liability under the CFAA.

Initially, many courts focused on whether a website's "terms of use" or "terms of service" policy prohibited Web-scraping activities, and, if so, whether the website user accessed the website in an "unauthorized" way (e.g., see EF Cultural Travel BV v. Zefer Corp. 2003; Southwest Airlines Co. v. Farechase, Inc. 2004; EarthCam, Inc. v. OxBlue, Inc. 2017). Since then, other courts have generally required some affirmative action on the website user to become a party to the terms of use/service, and, consequently, for the website's owners to deem such access unauthorized (see, e.g., Alan Ross Machinery Corp. v. Machinio Corp 2018). In 2016, the Ninth Circuit ruled that violating a website's terms of use/service alone cannot form the basis for liability under the CFAA (Facebook, Inc. v. Power Ventures, Inc. 2016).

However, a website's owner can revoke access to it and make access unauthorized by sending a cease and desist letter to a party crawling or scraping the website (Craigslist Inc. v. 3Taps Inc. 2013; Facebook, Inc. v. Power Ventures, Inc. 2016). Still, at least one court found a cease and desist letter alone as insufficient to hold the Web scraper liable under the CFAA (Ticketmaster L.L.C. v. Prestige Entertainment, Inc. 2018). In 2017, the District Court for the Northern District of California required a website owner to allow Web scraping on the owner's website despite a cease-and-desist letter and an IP block when the user accessed publicly available data (hiQ Labs, Inc. v. LinkedIn Corp. 2017). In September, 2019, the Ninth Circuit Court of Appeals upheld the lower court's decision in the hiQ case and noted that Web scraping publicly available data likely does not violate the CFAA (hiQ Labs, Inc. v. LinkedIn Corp 2019).

4.2 Breach of Contract

As we note above, legal professionals have pointed out that a website owner can effectively prevent programmatic access to a website by explicitly prohibiting such access in the website's terms of use/service. In addition to unauthorized access and use, failure to comply with these terms may lead to a "breach of contract" on the side of the website's user (Dreyer & Stockton, 2013). But to hold someone liable for violating the terms of use/service, the website user generally needs to enter an explicit agreement to comply with the policy (e.g., by clicking on a checkbox) (Alan Ross Machinery Corp. v. Machinio Corp 2018; Facebook, Inc. v. Power Ventures, Inc. 2016). Thus, simply prohibiting Web scraping on the website may not preclude someone from crawling it from a legal standpoint. Further, the website must establish that it incurred material damages due to the breach to its terms of use/service in order to succeed with a breach of contract claim.

In some jurisdictions, Web users have been enjoined or prevented from (or held liable for) Web-scraping activities due to a contract between the website and a third party who owns some or part of the content on the website. In such cases, Web scrapers have been held to the terms of the contract between the website and the third party generally because the website's terms of use/service require compliance with contracts between the website and third parties (e.g., QVC, Inc. v. Resultly, LLC 2016).

4.3 Copyrighted Material

Scraping and republishing data that the website owner owns and explicitly copyrights can lead to a copyright infringement case, especially when the party uses the scraped data for financial gain (Dreyer & Stockton, 2013). However, copyright law does not prevent one from collecting the data itself, nor does a website necessarily own the data on its website, particularly when users generate it (e.g., content from social media sites). Moreover, one cannot copyright ideas—only their specific form or representation. Thus, for example, one can use copyrighted data to summarize this data.

Also, one can still use copyrighted material on a limited scale under the "fair use" principle when one transforms the copyrighted material in a new or original way. However, this area remains a "grey area". While some courts have found that scrapers used scraped material according to the fair use principle (Kelly v. Arriba Soft Corp. 2003), others have concluded the opposite (Associated Press v. Meltwater U.S. Holdings, Inc. 2013). In the Meltwater case, the court indicated that using even a small percentage of scraped data—"as little as 4.5%"—could be enough to not fall under the fair use exception. However, a noteworthy aspect in the Meltwater case concerned the fact that Meltwater made the data available for purchase; if anyone could publicly access such data, the court may have reached a different conclusion.

Courts have also begun to consider whether authorization mechanisms or lack thereof constitute a license to copy scraped data. The robots.txt protocol constitutes one such mechanism. The instructions contained in a robot.txt file uploaded to a Web server can inform scrapers about any website restrictions or limitations on Web scraping, which includes website areas that they may access and expectations regarding crawl rate (Sellars, 2018). In Parker v. Yahoo (2008), the court suggested that a failure to include a robots.txt file with specific instructions could potentially create an implied license under copyright law. In Field v. Google, Inc. (2006), another court noted that the instructions that a website's robots.txt file provided prevented the website's owner from arguing that a scraper used the website's data in a way that infringed copyright. However, when a website's owner does not own the data on the website, the robots.txt file cannot constitute an implied license (Associated Press v. Meltwater U.S. Holdings, Inc. 2013).

4.4 Trespass to Chattels

If Web scraping overloads or damages a website or a Web server, then the person responsible for the damage may be liable under the "trespass to chattels" theory (Dreyer & Stockton, 2013). In 2000, a court awarded an injunction based on trespass against a Web scraper when the scraper prevented a website from using a small amount of server resources for other uses (eBay, Inc. v. Bidder's Edge, Inc. 2000). However, since then, courts have ruled that the damage needs to be material and easy to prove in court in order for the Web server's owner to be eligible for financial compensation (Intel Corp. v. Hamidi 2003). Since one often cannot easily prove such damage, Web scraping cases do not widely use the the trespass to chattels theory (Gold & Latonero, 2018).

4.5 Trade Secrets

One should not use Web scraping as a deliberate surveillance mechanism to try and reveal a competing organization's trade secrets. For example, some have accused Uber of using Web scraping to "spy" on competing companies and individual drivers (Rosenblatt, 2017). A trade secret refers to a legal concept that, in accordance with the United States Patent and Trademark Office (2020), constitutes a type of intellectual property that one can protect in court. According to the United States Patent and Trademark Office (2020) definition:

> *Trade secrets consist of information and can include a formula, pattern, compilation, program, device, method, technique or process. To meet the most common definition of a trade secret, it must be used in business, and give an opportunity to obtain an economic advantage over competitors who do not know or use it.*

Thus, the confidential aspects of a company's operations must meet the above definition and criteria in order for the law to legally protect them as a trade secret. If a certain aspect of a company's operations does not meet this definition (and yet the company still wants to keep it confidential), then the issue becomes related to organizational privacy, which we discuss in Section 5.

5 Ethics of Web Scraping

While both the courts and the legal literature have applied existing laws and legal theories to Web scraping, they have addressed Web scraping's ethics to a more limited degree. While many perspectives on ethics exist, ethical principles from the Association of Internet Researchers "Internet Research: Ethical Guidelines 3.0" (the IRE) best apply to Web scraping, and, hence, we adopt them in this paper (Association of Internet Researchers, 2019). The IRE promotes "ethical pluralism and cross-cultural awareness" and notes that European countries often take a deontological approach to ethics and privacy (i.e., they focus on means or duties rather than the end result), while the United States and United Kingdom focus on a more utilitarian approach (i.e., they emphasize the greatest good for the greatest number). With respect to real-world problems in the IS field, Mingers and Walsham (2010) recognize that one can use different ethical approaches to address various issues and advocate that researchers use discourse ethics, which synthesizes both deontological and utilitarian approaches. Discourse ethics posits that individuals affected by certain decisions debate the issues arising from these decisions in order to define universal ethical norms. The IRE suggests that one use a similar pluralistic approach in which one should acknowledge and explore a range of ethically defensible positions to a problem (e.g., deontological and utilitarian approaches, among others) while recognizing that some basic norms emerge among different ethical approaches.

Although applied differently, in both deontological and utilitarian approaches, researchers must generally consider the harms caused due to their actions. In addition to violating existing laws, Web scraping can result in unintended harm to others, such as the website's owners or customers. One can often not easily predict these harmful consequences (Light & McGrath, 2010). With that said, we discuss some possible harmful consequences from web Scraping in Sections 5.1 to 5.6.

5.1 Web Crawling Restrictions Provided

In the Web-scraping context, researchers need to first address whether a website includes a robots.txt file that prohibits them from conducting automated Web crawling. No legal or technical restrictions impel a Web user to follow the instructions that the robots.txt protocol provides. Still, social norms generally dictate that Web scrapers follow these instructions (Gold & Latonero, 2018). Various legitimate reasons (e.g., privacy concerns) can lead website creators to prohibit robots from automatically crawling their websites. Failure to obey these instructions can cause unintended harm to a website's owner and users.

5.2 Individual Privacy and Rights of Research Subjects

A research project that relies on data collected from a website may unintentionally compromise the privacy of individuals who participate in the activities that the website affords (Mason, 1986). For example, by matching the data collected from a website with other online and offline sources, a researcher could unintentionally reveal the identity of the individuals who created the data (Ives & Krotov, 2006). Many examples exist. For instance, after America Online (AOL) released approximately 500,000 search queries that its users submitted to its search engine in 2006, the Internet community quickly identified some users

due to “ego searches” that involved their own names and specific geographical locations even though the company replaced users’ names with random numbers. The findings derived from this dataset led to various ethical controversies. Some AOL members performed disturbing searches that indicated possible drug use, violence, and illegal pornographic materials.

Although website users may have no legal or technical protection for data shared on certain websites, they *expect* that websites somehow protect and keep such data private, which raises concerns, especially through the lens of deontological ethics. Broader and serious implications arise once a website publishes data, especially when “personal and sensitive data…could be used directly or indirectly against individuals in [certain] countries and political systems” (Association of Internet Researchers, 2019). For example, security and law enforcement officials could identify individuals through Web-scraping activities to prosecute criminal activity (Zuboff, 2015), which creates issues with a user’s assumptions about privacy and unjust rights violations. Additionally, Web scraping certain DNA ancestry databases could reveal users’ private genetic and health-related information, and various governments and organizations could use it to create a “biometric database useful for identifying nearly any American from a DNA sample” (Regalado, 2019).

Even if one does not violate individuals’ privacy when Web scraping, a website’s customers may not have consented to a third party’s using their data. Thus, using this data without consent violates research subjects’ deontological right to autonomy and equality (Buchanan, 2017). Given that they simply cannot obtain informed consent in certain Web-scraping projects, researchers must consider the steps necessary to protect confidentiality, including deleting certain identifiable information or “pseudonymiz[ing] their data separating keys from the actual data set” (Association of Internet Researchers, 2019). These privacy and rights violations can lead to serious consequences for a website’s owner given the heightened concern with online privacy in the light of the recent privacy scandals that involve such organizations as Facebook and Cambridge Analytica.

Further, users disclose information with the understanding that the receiver will use it in a certain context. When one uses it for purposes that users did not intend, additional issues arise regarding privacy and consent. For example, when booking a flight on a travel website, users typically know that the website may use the information to offer hotel or other travel-related accommodations. On the other hand, users would likely not expect a third party to use the data to make pricing decisions due to Web-scraping activities (Martin, 2015). Elsewhere, users understand that they need to accept advertisements in their browsers to use a social media service, but they do not always appreciate how certain parties might collect, use, and disclose their personal information through Web-scraping projects, which allows such parties to exploit and traffick personal data (Wigan & Clarke, 2013).

5.3 Discrimination and Bias

Information from Web-scraping activities can contribute to discriminatory practices, inferences drawn on preexisting biases, and prejudicial labeling. In addition to denying due process, decisions based on data that one obtained from Web scraping could lead to new forms of financial and social discrimination (Wigan & Clarke, 2013) or unfair profiling practices (Someh, Davern, Breidbach, & Shanks, 2019). Predictive algorithms based on prior data patterns may lead to learned prejudice built on institutionalized prejudice (e.g., in the college-admission process) (Martin, 2015). Businesses may target products and services to certain groups based on past behavior (limiting consumer choice) and charge different amounts for such goods and services based on gender, age, race, ethnicity, socioeconomic status, and so on (e.g., charging young men higher rates for car insurance) (Newell & Marabelli, 2015). Researchers should foresee and avoid such potential ways to use Web data.

5.4 Organizational Privacy

Just like individuals have the right to privacy, organizations also have the right to maintain certain operational aspects confidential (Mason, 1986). Automatic Web scraping can unintentionally reveal confidential information about an organization’s operations. For example, by automatically crawling and counting employment ads on an online recruitment website, one could approximate the website’s target audience, market share, and revenue. One could also reveal some details and possibly flaws in the way that the website stores data (Ives & Krotov, 2006). Instances in which an organization has suffered a data privacy breach, employed discriminatory practices lacked due process for users, and so on can damage its reputation, create legal issues, and lead to material financial losses (Markus, 2017).

5.5 Diminishing Value for the Organization

If an individual accesses a website not via the typical Web interface, then the person will avoid the advertisements that the website uses to monetize its content. Moreover, a Web-scraping project can lead one to create a data product (e.g., a report) that, without infringing on the copyright, makes it less likely for a customer to purchase a data product from the data's original owner. In other words, the data product that one creates with Web scraping's help can directly or indirectly compete the website's owner's business (Hirschey, 2014). Such activities may lead to financial losses to the website's owner or, at a minimum, an unfair distribution of value from data ownership (Mason, 1986).

On the other hand, some website owners partner with researchers in attempt to find ways to monetize these Web-scraping projects or for better brand recognition. Naturally, researchers also should carefully consider how such partnerships affect their research's independence (Association of Internet Researchers, 2019).

5.6 Data Quality and Impact on Decision-Making

Organizations and public officials often make strategic decisions based on data that they amass through Web-scraping activities, which can lead to faulty decision making due to Web data's veracity. Due to the anonymous, pluralistic, and "at-will" mechanisms that actors on the Web use to generate data on the Web, such data often lacks completeness, accuracy, and relevance (Clarke, 2016). Despite these issues, the need for consumer analytics can create destructive demand for such data and may encourage Web-scraping activities that collect and sell low-quality Web data (Martin, 2015). Further, a group that inappropriately uses Web data can negatively impact all individuals who use that data in the value chain, which magnifies the problem (Someh et al., 2019). To compound these issues, users often lack the ability to modify aggregated and shared data for accuracy—something that contributes to a likely exponential growth of what the data misrepresents (Someh et al., 2019). Some Internet users deliberately create and spread false information on the Web, which contributes to the problem about Web data quality even further and lowers the public's trust in any conclusions that one derives from the data.

Thus, organizations may base important decisions on haphazard, user-generated Web data that lacks quality, relevance, and compliance with strict academic or professional standards (Constantiou & Kallinikos, 2015). As a result, organizations that employ this data may make flawed decisions or suffer financial losses, and research that employs it may make flawed conclusions. Furthermore, the general public may treat low-quality Web data that they cannot easily (or at all) properly interpret as authoritative, which could allow organizations to manipulate consumer behavior and public opinion, limit consumer and voter choices, and create artificial markets and political agendas (Wigan & Clarke, 2013). These erroneous moves and decisions pose unintended consequences for entire economies and societies.

6 Legality and Ethics Framework for Web Scraping

Based on analyzing the literature that we present in this paper, we generate some questions that both researchers and practitioners need to address in order to make a Web-scraping project legal (Dreyer & Stockton, 2013; Hirschey, 2014; Snell & Menaldo, 2016) and ethical (Mason, 1986; Ives & Krotov, 2006; Buchanan, 2017):

- Does the website's terms of use/service explicitly prohibit Web crawling or scraping?
- Does the website explicitly copyright its data?
- Does the project involve illegal or fraudulent use of the data?
- Can crawling and scraping potentially cause material damage to the website or the Web server that hosts it?
- Has the website sent the user a cease and desist letter, blocked the user's IP address, or closed access to data in some other way?
- Does the website's robots.txt protocol significantly limit or prevent Web-scraping activities?
- Can the data obtained from the website compromise individual privacy, research subjects' rights, or non-discrimination principles?
- Can the data obtained from the website reveal confidential information about organizations affiliated with the website?

- Can the project that requires the Web data potentially diminish the value of the service that the website provides?
- Does the quality of the data obtained from the Web have the potential to lead to ill-informed decision making?

We derived these questions from the broad legal frameworks and ethical issues surrounding Web scraping that we discuss in Section 5. The legal frameworks or principles on which the framework builds include illegal access of data, illegal use of data, breach of contract, copyright, trespass to chattels, and trade secrets. The ethical dimension includes such concepts as Web crawling restrictions (imposed by a website's owner), individual privacy, rights of research subjects, organizational privacy, diminishing value for the organization, discrimination and bias, and data quality and decision making. We describe these ethical and legal concepts in more detail in Section 5.

We visually depict these questions, legal frameworks, and ethical concerns in Figure 2, which we call the legality and ethics framework for Web scraping. However, we draw no specific links between questions and legal and ethical elements since they significantly overlap. Moreover, the links are "loosely coupled" and, therefore, "non-propositional", a defining characteristic of a theoretical framework that does not claim to be a theory (Bacharach, 1989).

Figure 2. Legality and Ethics Framework for Web Scraping

A positive answer to any question in Figure 2 may suggest that the Web-scraping project can potentially result in lawsuits or ethical controversies. However, researchers may not need to halt a research project that potentially violates one or more principles that we discuss in this paper. For example, one can still use copyrighted data in accordance with the fair use principle. Even if terms of use/service prohibit crawling, one can still obtain permission to automatically collect data from the website's owner. Still, researchers behind the projects that involve a positive answer to any question in Figure 2 should reflect on how they will deal with the arising issues in order to avoid ethical controversies or even lawsuits. We provide an example study that relies on Web scraping and address potential concerns in the study based on Figure 2 in Appendix A.

We also note that the list of questions in Figure 2 may not exhaustively list all the legal and ethical controversies that a particular Web-scraping project can produce. The data available on the Web and the various tools, technologies, and applications related to this data continue to grow at a rapid pace. Thus, one would find it hard (or impossible) to predict all areas of possible future concern in relation to using

Web data (Light & McGrath, 2010). We do recommend that researchers always ask themselves the following two overreaching questions that targets both legal and ethical considerations:

- Can my actions in relation to Web data produce harm to individuals, organizations, or communities?
- What can we do to provide reasonable assurance that we do not unintentionally produce such harm?

In particular, academic researchers whose respective institutional review board or human subjects protection committee require them to use all possible mechanisms to protect their research subjects' privacy and wellbeing need to reflect on these issues and topics. We recommend that academic journals should also incorporate these questions as a formal requirement for their manuscript-submission process to provide an additional layer to protect human subjects that researchers use in their studies and the interests of other related stakeholders, such as academic journal publishers. We also hope that various academic and industry research organizations will incorporate these questions into their policies and procedures to protect human subjects, individual researchers, and organizations that conduct research from harm that may result from ethical and legal controversies. We believe that doing so can foster research that relies on data from the Web, which has become a vast repository of valuable datasets that span both the social and material realms.

7 Conclusion

The big data available on the Web presents researchers and practitioners with numerous lucrative opportunities. For researchers, these opportunities include leveraging this data to better understand various old and new social phenomena with better timeliness and precision. Practitioners can leverage this data to better understand their customers and develop data products that help with decisions and strategic planning. But leveraging big data from the Web presents both researchers and practitioners with big challenges as well. Apart from the need to learn and deploy new tools and technologies that can accommodate big data, researchers and practitioners who intend to use Web scraping in their research projects need to comply with various legal and ethical requirements. Unfortunately, due to the Web-scraping phenomenon's relative novelty, Web scraping's legality and ethics remain grey areas. In this paper, we reflect on some legal and ethical issues that surround Web scraping. We formulate some specific questions that researchers who employ Web scraping need to address. A negative answer to all these questions does not necessarily give a clearance to proceed with a research project. Rather, researchers should use the questions as a starting point to reflect on the legality and ethics of a research project that relies on Web scraping to acquire data. Researchers should always focus on identifying potential harm to individuals and organizations that result from their projects and on implementing reasonable precautions to prevent this harm from occurring. Finally, we stress that the information that we provide in this paper does not, and we do not intend for it, to constitute legal advice. We urge both researchers and practitioners to seek qualified legal help when in doubt regarding their Web-scraping projects' legality.

References

Alan Ross Machinery Corp. v. Machinio Corp, 2018 WL 6018603 (N.D. Ill. Nov. 16, 2018).

Associated Press v. Meltwater U.S. Holdings, Inc. 931 F. Supp. 2d 537 (S.D.NY 2013).

Association of Internet Researchers. (2019). *Internet research: Ethical guidelines 3.0.* Retrieved from https://aoir.org/reports/ethics3.pdf

Bacharach, S. B. (1989). Organizational theories: Some criteria for evaluation. *Academy of Management* Review, *14*(4), 496-515.

Basoglu, K. A., & White, C. E., Jr. (2015). Inline XBRL versus XBRL for SEC reporting. *Journal of Emerging Technologies in Accounting*, *12*(1), 189-199.

Boeing, G., & Waddell, P. (2017). New insights into rental housing markets across the United States: Web scraping and analyzing craigslist rental listings. *Journal of Planning Education and Research*, *37*(4), 457-476.

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.

Chaulagain, R. S., Pandey, S., Basnet, S. R., & Shakya, S. (2017). Cloud based Web scraping for big data applications. In *Proceedings of the IEEE International Conference on Smart Cloud.*

Cisco. (2016). *Cisco visual networking index: Forecast and methodology, 2014-2019.* Retrieved from https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

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

Computer Fraud and Abuse Act of 1984, 18 U.S.C. § 1030 (2008).

Constantiou, I. D., & Kallinikos, J. (2015). New games, new rules: Big data and the changing context of strategy. *Journal of Information Technology*, *30*(1), 44-57.

Craigslist Inc. v. 3Taps Inc., 942 F.Supp 2d 962 (N.D. Cal. 2013).

DHI Group. (2017). *About DHI.* Retrieved from https://dhigroupinc.com/about-dhi/default.aspx

Dreyer, A. J., & Stockton, J. (2013). Internet “data scraping”: A primer for counseling clients. *New York Law Journal*. Retrieved from https://www.law.com/newyorklawjournal/almID/1202610687621

EarthCam, Inc. v. OxBlue, 703 Fed. Appx. 803 (11th Cir. 2017).

eBay, Inc. v. Bidder’s Edge, Inc., 100 F. Supp. 2d 1058 (N.D. Cal. 2000).

EF Cultural Travel BV v. Zefer Corp., 318 F.3d 58 (1st Cir. 2003).

Facebook, Inc. v. Power Ventures, Inc. et. al., 844 F.3d 1058 (9th Cir. 2016).

Field v. Google Inc., 412 F. Supp. 2d 1066 (D. Nev 2006).

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

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Martinez-Hernandez, V., Sanchez, V., & Perez-Meana, H. (2018). A web scraping methodology for bypassing Twitter API restrictions. *arXiv*. Retrieved from https://arxiv.org/abs/1803.09875

hiQ Labs, Inc. v, LinkedIn Corp., 273 F.Supp. 3d 1099 (N.D. Cal. 2017).

hiQ Labs, Inc. v. LinkedIn Corp., 938 F. 3d 985 (9th Cir. 2019).

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

IBM. (2018). *The four V’s of big data.* Retrieved from http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Intel Corp. v. Hamidi, 71 P. 3d 296 (2003).

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*, 593-611.

Ives, B., Palese, B., & Rodriguez, J. A. (2016). Enhancing customer service through the Internet of things and digital data streams. *MIS Quarterly Executive*, *15*(4), 279-297.

Kelly v. Arriba Soft Corp., 336 F.3d 811 (9th Cir. 2003).

Kirkpatrick, K. (2015). Putting the data science into journalism. *Communications of the ACM*, *58(5),* 15-17.

Krotov, V., & Silva, L. (2018). *Legality and ethics of Web scraping.* In *Proceedings of the 24th Americas Conference on Information Systems.*

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.

Light, B., & McGrath, K. (2010). Ethics and social networking sites: A disclosive analysis of Facebook. *Information Technology & People*, *23*(4), 290-311.

Markus, M. (2017). Datification, organizational strategy, and IS research: What's the score? *Journal of Strategic Information Systems*, *26*(3), 233-241.

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.

Mingers, J., & Walsham, G. (2010). Towards ethical information systems: The contribution of discourse ethics*. MIS Quarterly*, *34*(4), 833-854.

Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2015). Automated data collection with R: A practical guide to Web scraping and text mining. Chichester, UK: John Wiley & Sons.

Neumann, M., Steinberg, J., & Schaer, P. (2017). Web-Scraping for non-programmers: Introducing OXPath for digital library metadata harvesting. *Code4Lib Journal*, *38*.

Newell, S., & Marabelli, M. (2015). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of "datification". *The Journal of Strategic Information Systems*, 24(1), 3-14.

Parker v. Yahoo!, Inc. No. 07-2757, 2008 WL 4410095 (E.D. Pa. September 25, 2008).

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226-1227.

QVC, Inc. v. Resultly, LLC, 159 F. Supp. 3d 576 (E.D. Pa 2016).

Regalado, A. (2019). The DNA database used to find the Golden State Killer is a national security leak waiting to happen. *MIT Technology Review.* Retrieved from https://www.technologyreview.com/2019/10/30/132142/dna-database-gedmatch-golden-state-killer-security-risk-hack/

Rosenblatt J. (2017). Uber data-scraping, surveillance detailed by ex-manager. *Bloomberg*. Retrieved from https://www.bloomberg.com/news/articles/2017-12-15/uber-data-scraping-surveillance-detailed-in-ex-manager-s-letter

Sagina, I. J. (2018). The ultimate list of Web scraping tools and software. *KDnuggets*. Retrieved from https://www.kdnuggets.com/2018/07/ultimate-list-web-scraping-tools-software.html

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

Snell, J., & Menaldo, N. (2016). Web scraping in an era of big data 2.0. Bloomberg BNA. Retrieved from https://www.perkinscoie.com/images/content/1/5/v2/156775/Snell-web-scraping-BNAI.pdf

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*, 718-747.

Southwest Airlines Co. v. Farechase, Inc. 318 F. Supp. 2d 435 (N.D. Tex. Mar. 19, 2004).

Ticketmaster L.L.C. v. Prestige Entertainment, Inc. et al., No. 17-cv-07232, 2018 WL 654410 (C.D. Cal. Jan. 31, 2018).

Ujwal, B. V. S., Gaind, B., Kundu, A., Holla, A., & Rungta, M. (2017). Classification-based adaptive Web scraper. In *Proceedings of the 16th IEEE International Conference on Machine Learning and Applications.*

United States Patent and Trademark Office. (2020). *Trade secret policy*. Retrieved from https://www.uspto.gov/ip-policy/trade-secret-policy

Watson, H. J. (2014). Big data analytics: Concepts, technologies, and applications. *Communications of the Association for Information Systems*, *34*, 1247-1268.

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

Zamora, A. (2019). Making room for big data: Web scraping and an affirmative right to access publicly available information online. *The Journal of Business, Entrepreneurship, & The Law*, *12*(1), 203-226.

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, *30*(1), 75-89.

Appendix A: The Case of Dice.com

In this appendix, we apply the legality and ethics framework for Web scraping that we propose to Dice (www.dice.com)—a leading specialized recruitment website for IT and engineering professionals (DHI Group, 2017). In doing so, we focus on answering the following research question:

RQ: What skills or requirements do employers most demand a systems analyst to have?

We can potentially answer this question using the data available on Dice. Historically, the website has specialized in job search and recruitment solutions for IT and engineering workforce (DHI Group, 2017). At a given moment, the website contains close to 80,000 active job listings that one can publicly access, and it has over 2.1 million resumes (not in public access). The website also reports that it has more than two million unique visitors every month.

At its most basic level, we can view Dice.com as an online bulletin board where potential job seekers can post their resumes and recruiters representing various companies can post their job ads. Job seekers can then apply to posted jobs, while recruiters can browse through the resumes of IT professionals who use the website. Dice.com offers several other research and data solutions related to IT recruitment for companies in finance, energy, healthcare, and hospitality (e.g. security clearance of job applicants). The company also publishes research in relation to IT labor and recruitment.

Phase 1: Website Analysis

To analyze Dice, we first studied the API manual that it provided (no longer available). From this manual, we could see that the website allowed the public to use one domain to access job data: "service.dice.com". One can obtain the data on the website using the API in three formats: HTML, XML, and JSON. One can query the website's database of job listings by supplying the URL devoted to one of these data formats with parameters. Among other things, these parameters allow one to narrow down the search to specific keywords or pre-format the resulting output from the Web server. We provide sample output from an API search query that we formed to retrieve job listings related to the role of a "systems analyst" in Appendix B.

As one can see from the output in JSON format (see Appendix B), the API does not allow one to achieve actual job descriptions. The API allows only URLs that point to actual job descriptions. We opened one of these URLs in a browser (Google Chrome) for further analysis. In Google Chrome, one can use the "inspect" feature to analyze a job description page's underlying code in order to determine which elements contain the needed job description. One needs this code to retrieve the actual job description for each job listing with the API using an R script (which we discuss in more detail below).

Reflecting on Legality and Ethics of Crawling and Scraping Dice.com

Before commencing the actual data-collection process based on the technical analysis that we describe above, we reflect on data quality and legality and ethics of this Web data-collection project.

First, we concluded that Dice.com provides high-quality data on job descriptions. Technology and hiring professionals mostly posted job ads, which the website administrators moderated further. Thus, we believed that we would not likely see glaring problems with data quality that would possibly lead to flawed decision making.

Subsequently, we examined the terms of use agreement for the website. These terms explicitly allowed one to access data using the API created by the company created. Developers can use this API at no charge. One can use the API from Dice to programmatically search through jobs ads and, if necessary, download this information. Thus, we concluded that the company encouraged the user community to automatically crawl and scrape the website. Programmatically retrieving job descriptions outside the data that the official API returns raises a slight ethical dilemma. Most websites, including Dice.com, have an item in their terms of use that prohibits one from accessing website data using unauthorized means. Thus, we needed to reflect on whether we accessed the job descriptions in an authorized way. After some deliberation, we concluded that we would not do anything illegal or unethical by going outside the data that the API returned (see Appendix B). First, the API returned URLs that pointed to actual job descriptions, which meant that the organization did not want to keep those URLs secret. Second, the URLs pointed to publicly available pages. One does not even need an account with Dice.com to view

these pages. Thus, we concluded that we could go ahead and retrieve the job descriptions that these URLs returned via the API.

We did not find any restrictions for automatically crawling the website areas that contained job ads in the robots.txt file either. Thus, we concluded that we could go ahead and develop an R script that automatically crawled the pages and downloaded data.

The rate at which we retrieved job descriptions from the website (approximately one job description every one or two seconds, which we determined through some "trial" runs using smaller sets) and the relatively small dataset (1105 job ads) did not make it likely that automatically crawling and retrieving the data we needed would damage or slow down the website's servers.

We also saw that the website contained a statement that it copyrighted the data it provided. Since we did not intent to republish the data we collected verbatim (we only intended to summarize the most frequent keywords in those job ads at a high level), we concluded that we would comply with the fair use principle in using this data. We also determined that, if one somehow aggregated the data we collected at the company level, it could expose certain aspects of the operations of the companies that posted these job ads to Dice. Again, since we did not intend to republish this data verbatim or group it using company names, we felt like we properly respected organizational privacy and trade secrets.

Overall, we believed that we used the data for a legitimate, non-fraudulent purpose (i.e., research). We did not believe that our results would somehow diminish the website's data. If anything, such research could provide additional exposure to the company's brand and provide a case for the data's quality and usefulness.

Thus, we concluded that we used the website's data in a legal and ethical way.

Phase 2: Website Crawling

In this phase, we first developed and debugged an R script to crawl the website and download the data related to systems analyst job listings. Next, we ran the script with some degree of human supervision to retrieve the data. This phase produces a data frame (a popular data structure in R similar to a table in Excel) filled with data and meta-data related to systems analyst job listings.

We chose R as a Web-scraping tool for several reasons. First, one can use R without cost. Using some ready-made scraping tools would probably involve paying a hefty subscription fee. Second, R contains several useful libraries (e.g., rvest) to simulate Web sessions and parse data in various formats. Third, one can fine-tune R code to do a Web-scraping task of any complexity or granularity.

We first applied the developed R script to a moderately sized data set that comprised 1,104 observations (23 webpages that contained job descriptions). It took approximately 19 minutes to download data related to 1,104 jobs on an office computer with the following configuration:

- Processor: Intel® Core™ i5-4590 CPU @ 3.30 GHz
- RAM: 16GB
- Upload/Download Speed: ~100 Mbps

It took approximately 12 hours to download data for 44,889 job listings that contained words "systems" and "analyst" (not necessarily as a single phrase) using the computer with the configuration above.

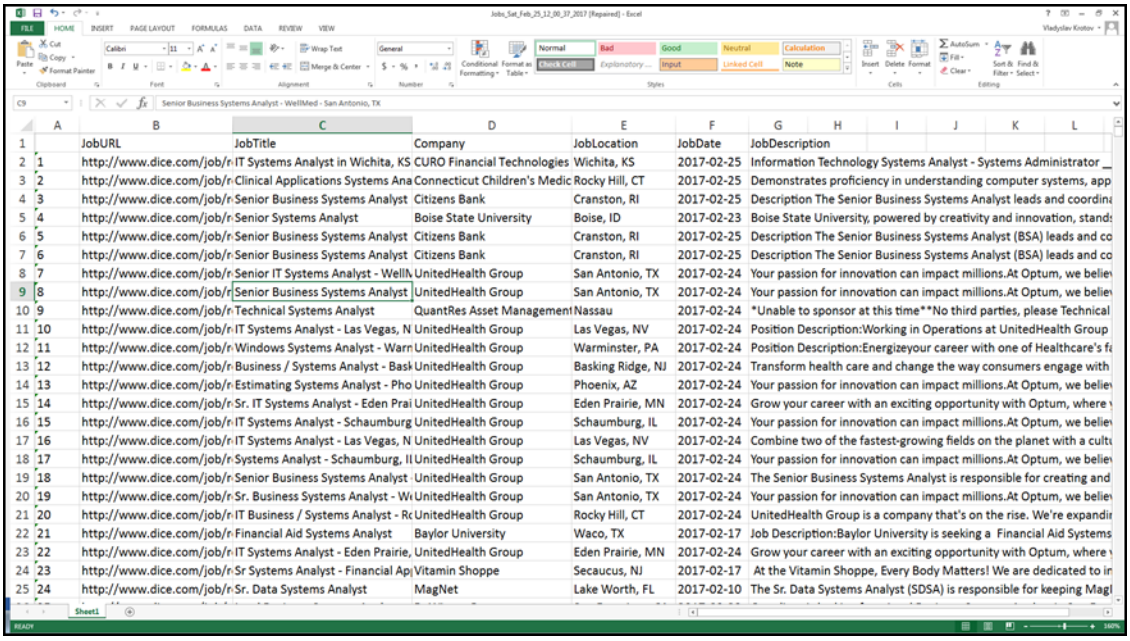
We also observed that we queried the server to obtain approximately one job description per second. We concluded that such crawling rate would not likely to overload the company's Web servers and prevent other users from accessing its resources.

Phase 3: Data Organization

Once we saved all the scraped data in the job_table data frame, we used R to further process and organize it in a Microsoft Excel sheet. Using the "xlsx" R package function, we saved the entire data frame into one single Excel file. We used Excel files to store data due to the format's universality and popularity. Researchers often use Excel to manually analyze data. Most people view and analyze data in Excel quite comfortably. Moreover, one can easily import Excel files into virtually any other software (including R and NVivo).

The file name contained the additional metadata: date and time. We added this data so that we could determine when we conducted a particular scraping job by simply looking at its file. Some research projects may require researchers to scrape data on multiple dates to come up with a substantially large data sample.

Figure A1 shows how the Excel sheet organized the data. The file in Figure A1 contains information about systems analyst jobs listed on Dice. Each row corresponds to a particular job listing, which clearly delineates job ads. Each column represents a particular attribute of each job listing. The specific columns include the URL used to obtain a job description (JobURL), the title of the job listed (JobTitle), the name of the company that posted the job (Company), when the company posted the job (JobDate), and job listing's full description (JobDescription).



	A	JobURL	JobTitle	Company	JobLocation	JobDate	JobDescription
2	1	http://www.dice.com/job/r	IT Systems Analyst in Wichita, KS	CURO Financial Technologies	Wichita, KS	2017-02-25	Information Technology Systems Analyst - Systems Administrator
3	2	http://www.dice.com/job/r	Clinical Applications Systems Ana	Connecticut Children's Medic	Rocky Hill, CT	2017-02-25	Demonstrates proficiency in understanding computer systems, app
4	3	http://www.dice.com/job/r	Senior Business Systems Analyst	Citizens Bank	Cranston, RI	2017-02-25	Description The Senior Business Systems Analyst leads and coordina
5	4	http://www.dice.com/job/r	Senior Systems Analyst	Boise State University	Boise, ID	2017-02-23	Boise State University, powered by creativity and innovation, stands
6	5	http://www.dice.com/job/r	Senior Business Systems Analyst	Citizens Bank	Cranston, RI	2017-02-25	Description The Senior Business Systems Analyst (BSA) leads and co
7	6	http://www.dice.com/job/r	Senior Business Systems Analyst	Citizens Bank	Cranston, RI	2017-02-25	Description The Senior Business Systems Analyst (BSA) leads and co
8	7	http://www.dice.com/job/r	Senior IT Systems Analyst - WellM	UnitedHealth Group	San Antonio, TX	2017-02-24	Your passion for innovation can impact millions.At Optum, we belie
9	8	http://www.dice.com/job/r	Senior Business Systems Analyst	UnitedHealth Group	San Antonio, TX	2017-02-24	Your passion for innovation can impact millions.At Optum, we belie
10	9	http://www.dice.com/job/r	Technical Systems Analyst	QuantRes Asset Management	Nassau	2017-02-24	*Unable to sponsor at this time**No third parties, please Technical
11	10	http://www.dice.com/job/r	IT Systems Analyst - Las Vegas, N	UnitedHealth Group	Las Vegas, NV	2017-02-24	Position Description:Working in Operations at UnitedHealth Group
12	11	http://www.dice.com/job/r	Windows Systems Analyst - Warn	UnitedHealth Group	Warminster, PA	2017-02-24	Position Description:Energizeyour career with one of Healthcare's f
13	12	http://www.dice.com/job/r	Business / Systems Analyst - Bask	UnitedHealth Group	Basking Ridge, NJ	2017-02-24	Transform health care and change the way consumers engage with
14	13	http://www.dice.com/job/r	Estimating Systems Analyst - Pho	UnitedHealth Group	Phoenix, AZ	2017-02-24	Your passion for innovation can impact millions.At Optum, we belie
15	14	http://www.dice.com/job/r	Sr. IT Systems Analyst - Eden Prai	UnitedHealth Group	Eden Prairie, MN	2017-02-24	Grow your career with an exciting opportunity with Optum, where
16	15	http://www.dice.com/job/r	IT Systems Analyst - Schaumburg	UnitedHealth Group	Schaumburg, IL	2017-02-24	Your passion for innovation can impact millions.At Optum, we belie
17	16	http://www.dice.com/job/r	IT Systems Analyst - Las Vegas, N	UnitedHealth Group	Las Vegas, NV	2017-02-24	Combine two of the fastest-growing fields on the planet with a cult
18	17	http://www.dice.com/job/r	Systems Analyst - Schaumburg, Il	UnitedHealth Group	Schaumburg, IL	2017-02-24	Your passion for innovation can impact millions.At Optum, we belie
19	18	http://www.dice.com/job/r	Senior Business Systems Analyst	UnitedHealth Group	San Antonio, TX	2017-02-24	The Senior Business Systems Analyst is responsible for creating and
20	19	http://www.dice.com/job/r	Sr. Business Systems Analyst - W	UnitedHealth Group	San Antonio, TX	2017-02-24	Your passion for innovation can impact millions.At Optum, we belie
21	20	http://www.dice.com/job/r	IT Business / Systems Analyst - Rc	UnitedHealth Group	Rocky Hill, CT	2017-02-24	UnitedHealth Group is a company that's on the rise. We're expandi
22	21	http://www.dice.com/job/r	Financial Aid Systems Analyst	Baylor University	Waco, TX	2017-02-17	Job Description:Baylor University is seeking a Financial Aid Systems
23	22	http://www.dice.com/job/r	IT Systems Analyst - Eden Prairie,	UnitedHealth Group	Eden Prairie, MN	2017-02-24	Grow your career with an exciting opportunity with Optum, where
24	23	http://www.dice.com/job/r	Sr Systems Analyst - Financial Ap	Vitamin Shoppe	Secaucus, NJ	2017-02-17	At the Vitamin Shoppe, Every Body Matters! We are dedicated to in
25	24	http://www.dice.com/job/r	Sr. Data Systems Analyst	MagNet	Lake Worth, FL	2017-02-10	The Sr. Data Systems Analyst (SDSA) is responsible for keeping Magl

Figure A1. Excel File Content

In addition to having one single Excel file that contained all scraped job listings, the script produced a separate Excel file for each page of job listings that we scraped. By default, the API returned information about 50 job listings at a time. Thus, separate Excel files each stored 50 job listings as well. These saved pages collectively duplicated the information in the main Excel sheet. We used such duplication to avoid complete data loss during the scraping process. Obtaining a dataset that contains thousands of observations may take a computer hours, or even days, of non-stop work. If script execution halts (e.g., the Web server becomes unavailable, the computer runs out of memory or freezes, a power outage occurs, etc.), researchers still have access to the data that they collected prior to the failure point. The script saves this data as Excel files up to the point when it stopped for whatever reason. If the script successfully executes, it saves the entire dataset in the main time-stamped Excel file (e.g., Jobs_Sat_Feb_25_12_00_37_2017.xlsx). The file name means that the script saved the data on 25 February, 2017 at 12:00:37 pm. We provide the entire file structure that we produced after running the Web-scraping task in Figure A2.

We stored these files on a password-protected computer that the first author owned. Thus, we concluded that we took adequate measures to protect what might constitute copyrighted data from unauthorized access. We agreed that we would not share the data with anyone else.

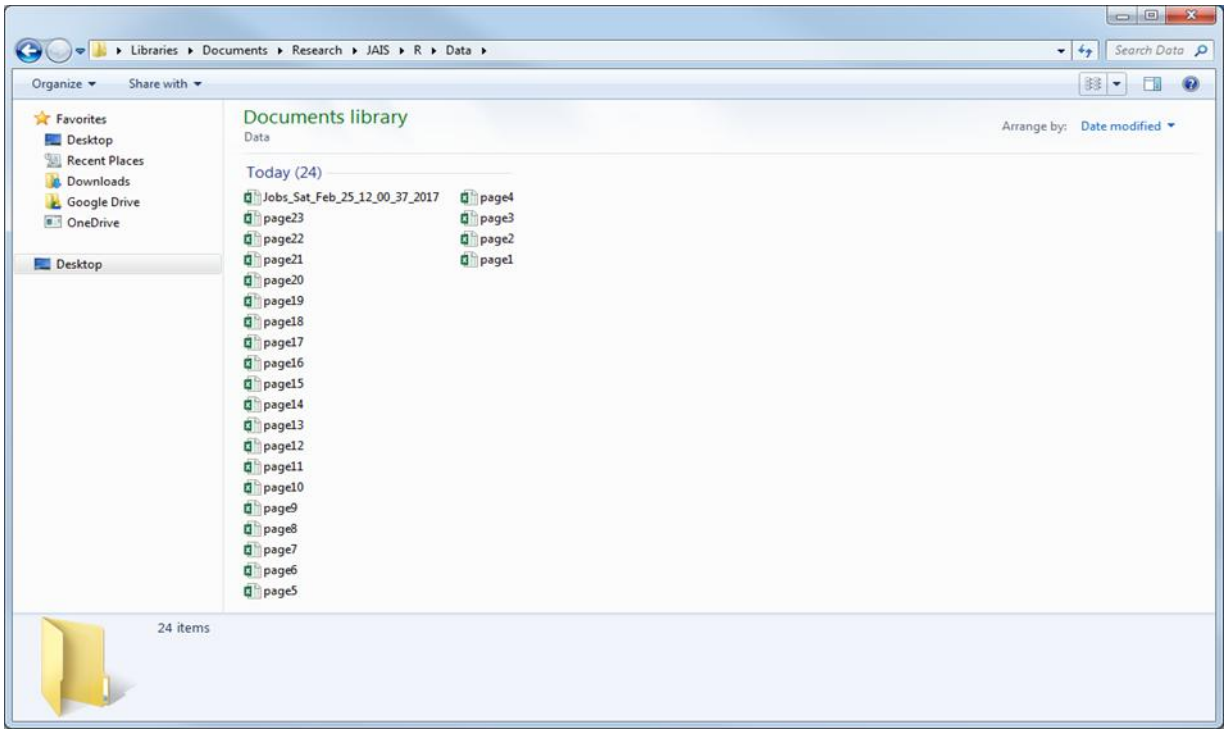


Figure A2. File Structure Created

What Happened Next

Some interesting developments related to the legality and ethics of the Web-scraping project that we describe in this appendix occurred several months after we executed the data-collection task. The first author developed an R package that could automatically crawl and retrieve data from Dice.com based on supplied parameters to stimulate further research in this area, save other researchers' time, and valuably extend the utilities available for the website. The author viewed his efforts as doing something valuable to the broader community using his time and expertise at no charge.

The first author eventually published the package on CRAN—a leading peer-reviewed repository that comprises user-provided R extensions. The author also wrote a manual on how to use this package and advertised its availability to various Web communities, such as the ISWorld mailing list. The package started gaining popularity. Several researchers approached the package's author for additional instructions and clarifications on how to use it.

A few months after the author released the R package, several individuals who used it notified the author that it no longer worked. The author quickly determined that the R code in the package had no errors. Instead, he determined that the API that Dice supplied and that the package used no longer responded. The author wrote an email to Dice asking why the API no longer worked. He also encouraged the package's users to make similar inquiries with Dice. At first, Dice sent standard "template" responses to each inquiry, such as "Thank you for your inquiry, if you are interested in any of our data products please visit dice.com". Eventually, the author received a more detailed response. The Dice representative stated that the company no longer supported the API. The representative added that the company intended its own internal developers to use the API rather than the public and that anyone interested in Dice data should use the main Web interface or subscribe to relevant data products.

At first, the author wanted to modify the package so that it could obtain data directly from the website by bypassing the APIs. However, he then interpreted the situation as Dice imposing a restriction on crawling and scraping its data. Thus, he decided that it would be unethical and, perhaps, illegal for him to continue developing tools for scraping data from Dice. He found this development rather disappointing as he viewed himself as a volunteer who contributed his time and expertise at no charge for making the website more useful to the research community.

After we recently examined the website's robots.txt file again, we confirmed that the author correctly identified the situation. We found that the website did not want automated Web crawlers to programmatically access its job listings. Indeed, the website possibly updated the file to make sure potential users access job listings on the website's main entryway. Thus, the author decided to abandon the package. He still hopes that one day he can modify the R code that it works with some other website that contains IT job listings in order to stimulate further research on in-demand IT skills and competencies.

Appendix B: JSON Output from Dice.com

The output below is generated using the following URL being sent to the server:

http://service.dice.com/api/rest/jobsearch/v1/simple.json?text=%22systems%20analyst%22

{"count":1105,"firstDocument":1,"lastDocument":50,"nextUrl":"/api/rest/jobsearch/v1/simple.json?areacode
\u003d\u0026country\u003d\u0026state\u003d\u0026skill\u003d\u0026city\u003d\u0026text\u003d%22sy
stems+analyst%22\u0026ip\u003d\u0026diceid\u003d\u0026page\u003d2","resultItemList":[{"detailUrl":"ht
tp://www.dice.com/job/result/RTX16c2f4/it-systems-analyst?src\u003d19","jobTitle":"IT Systems Analyst in
Wichita, KS at CURO Financial Technologies Corp","company":"CURO Financial Technologies
Corp","location":"Wichita, KS","date":"2017-02-
25"},{"detailUrl":"http://www.dice.com/job/result/appblok/146_171500?src\u003d19","jobTitle":"Senior
Business Systems Analyst","company":"Citizens Bank","location":"Cranston, RI","date":"2017-02-
25"},{"detailUrl":"http://www.dice.com/job/result/appblok/146_171502?src\u003d19","jobTitle":"Senior
Business Systems Analyst","company":"Citizens Bank","location":"Cranston, RI","date":"2017-02-
25"},{"detailUrl":"http://www.dice.com/job/result/appblok/2151_5000178091506?src\u003d19","jobTitle":"Cl
inical Applications Systems Analyst - Ambulatory","company":"Connecticut Children\u0027s Medical
Center","location":"Rocky Hill, CT","date":"2017-02-
25"},{"detailUrl":"http://www.dice.com/job/result/appblok/146_171501?src\u003d19","jobTitle":"Senior
Business Systems Analyst","company":"Citizens Bank","location":"Cranston, RI","date":"2017-02-
25"},{"detailUrl":"http://www.dice.com/job/result/RTX1be463/4891134?src\u003d19","jobTitle":"Senior
Systems Analyst","company":"Boise State University","location":"Boise, ID","date":"2017-02-
23"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/696317?src\u003d19","jobTitle":"Sr. IT Systems
Analyst - Eden Prairie, MN","company":"UnitedHealth Group","location":"Eden Prairie, MN","date":"2017-
02-24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/695890?src\u003d19","jobTitle":"Senior IT
Systems Analyst - WellMed - San Antonio, TX","company":"UnitedHealth Group","location":"San Antonio,
TX","date":"2017-02-
24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/700495?src\u003d19","jobTitle":"Sr. Business
Systems Analyst - WellMed - San Antonio, TX","company":"UnitedHealth Group","location":"San Antonio,
TX","date":"2017-02-
24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/700494?src\u003d19","jobTitle":"Estimating
Systems Analyst - Phoenix, AZ or Eden Prairie, MN","company":"UnitedHealth Group","location":"Phoenix,
AZ","date":"2017-02-
24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/701950?src\u003d19","jobTitle":"Senior Business
Systems Analyst - WellMed - San Antonio, TX","company":"UnitedHealth Group","location":"San Antonio,
TX","date":"2017-02-
24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/702044?src\u003d19","jobTitle":"Business /
Systems Analyst - Basking Ridge, NJ","company":"UnitedHealth Group","location":"Basking Ridge,
NJ","date":"2017-02-
24"},{"detailUrl":"http://www.dice.com/job/result/10449280/844229?src\u003d19","jobTitle":"Technical
Systems Analyst","company":"QuantRes Asset Management","location":"Nassau","date":"2017-02-
24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/701249?src\u003d19","jobTitle":"IT Systems
Analyst - Las Vegas, NV","company":"UnitedHealth Group","location":"Las Vegas, NV","date":"2017-02-
24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/696804?src\u003d19","jobTitle":"IT Systems
Analyst - Schaumburg, IL","company":"UnitedHealth Group","location":"Schaumburg, IL","date":"2017-02-
24"},{"detailUrl":"http://www.dice.com/job/result/jobtblok/96871132?src\u003d19","jobTitle":"Financial Aid
Systems Analyst","company":"Baylor University","location":"Waco, TX","date":"2017-02-
17"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/696554?src\u003d19","jobTitle":"Senior Business
Systems Analyst - WellMed - San Antonio, TX","company":"UnitedHealth Group","location":"San Antonio,
TX","date":"2017-02-
24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/699843?src\u003d19","jobTitle":"Windows
Systems Analyst - Warminster, PA","company":"UnitedHealth Group","location":"Warminster,
PA","date":"2017-02-
24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/692098?src\u003d19","jobTitle":"IT Systems
Analyst - Eden Prairie, MN","company":"UnitedHealth Group","location":"Eden Prairie, MN","date":"2017-

02-24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/663371?src\u003d19","jobTitle":"IT Systems Analyst - Las Vegas, NV","company":"UnitedHealth Group","location":"Las Vegas, NV","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/700855?src\u003d19","jobTitle":"IT Business / Systems Analyst - Rocky Hill, CT","company":"UnitedHealth Group","location":"Rocky Hill, CT","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/702997?src\u003d19","jobTitle":"Systems Analyst - Schaumburg, IL","company":"UnitedHealth Group","location":"Schaumburg, IL","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/10127643A/9698?src\u003d19","jobTitle":"Clinical Systems Analyst II - Ambulatory","company":"Allscripts","location":"Whittier, CA","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/RTX15c476/853002?src\u003d19","jobTitle":"Sr. Data Systems Analyst","company":"MagNet","location":"Lake Worth, FL","date":"2017-02-10"},{"detailUrl":"http://www.dice.com/job/result/10231917/4853133?src\u003d19","jobTitle":"Systems Analyst","company":"Great River Energy","location":"Elk River, MN","date":"2017-02-23"},{"detailUrl":"http://www.dice.com/job/result/RTL159177/830108?src\u003d19","jobTitle":"EMS Systems Analyst","company":"Associated Electric Cooperative, Inc.","location":"Springfield, MO","date":"2017-02-22"},{"detailUrl":"http://www.dice.com/job/result/10127643A/9677?src\u003d19","jobTitle":"Sr Clinical Systems Analyst","company":"Allscripts","location":"Whittier, CA","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/10127643A/9657?src\u003d19","jobTitle":"MS Systems Analyst-Clinical Applications","company":"Allscripts","location":"Whittier, CA","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/10436270/890183?src\u003d19","jobTitle":"Lead Business Systems Analyst","company":"DeWinter Group","location":"San Francisco, CA","date":"2017-02-23"},{"detailUrl":"http://www.dice.com/job/result/jobeblok/957965?src\u003d19","jobTitle":"Business Systems Analyst 3 (7584U) #22697","company":"University of California, Berkeley","location":"Berkeley, CA","date":"2017-02-03"},{"detailUrl":"http://www.dice.com/job/result/10116872/885080?src\u003d19","jobTitle":"Sr Systems Analyst - Financial Applications","company":"Vitamin Shoppe","location":"Secaucus, NJ","date":"2017-02-17"},{"detailUrl":"http://www.dice.com/job/result/10415537/309?src\u003d19","jobTitle":"Technical Business Systems Analyst","company":"Sila Solutions Group","location":"Shelton, CT","date":"2017-02-23"},{"detailUrl":"http://www.dice.com/job/result/10210023D/836527?src\u003d19","jobTitle":"Sr. Business Systems Analyst - Business Intelligence","company":"Crescent Solutions Inc","location":"Irvine, CA","date":"2017-02-06"},{"detailUrl":"http://www.dice.com/job/result/10519627/126207?src\u003d19","jobTitle":"Jr. Systems Analyst","company":"4M Research","location":"Huntsville, AL","date":"2017-01-26"},{"detailUrl":"http://www.dice.com/job/result/10428868/RT-103082?src\u003d19","jobTitle":"HR-Systems Analyst-Workday-Long Term-CTH-Pittsburgh-Immediate","company":"Accion Labs","location":"Pittsburgh, PA","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/cxelisen/261194?src\u003d19","jobTitle":"Business Systems Analyst","company":"Eliassen Group","location":"Boston, MA","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/10235494/880449?src\u003d19","jobTitle":"Sr. Systems Analyst","company":"Stellargy Services, LLC","location":"Reston, VA","date":"2017-02-15"},{"detailUrl":"http://www.dice.com/job/result/gatpa001/137896?src\u003d19","jobTitle":"Applications Systems Analyst/Programmer (Senior)","company":"Mastech","location":"Washington, DC","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/10124935/SP944956511375582?src\u003d19","jobTitle":"Senior Business Systems Analyst","company":"BIAS Corporation","location":"Atlanta, GA","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/10204742/JO170111043-869?src\u003d19","jobTitle":"Sr. Systems Analyst","company":"Strategic IT Staffing","location":"Reston, VA","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/10124488/886260?src\u003d19","jobTitle":"Business Systems Analyst","company":"Synergy Computer Solutions","location":"Mettawa, IL","date":"2017-02-20"},{"detailUrl":"http://www.dice.com/job/result/accperny/ERPEPIC-BB?src\u003d19","jobTitle":"ERP Systems Analyst - Epicor","company":"Access Staffing","location":"Bay Shore, NY","date":"2017-02-16"},{"detailUrl":"http://www.dice.com/job/result/cxtcml/17-00324?src\u003d19","jobTitle":"HRIS Business Systems Analyst","company":"Computer Merchant, Ltd., The","location":"Mentor, OH","date":"2017-02-07"},{"detailUrl":"http://www.dice.com/job/result/10124935/CC31226521679157?src\u003d19","jobTitle":"BA/Systems Analyst","company":"BIAS Corporation","location":"Brentwood, TN","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/10117123/890430?src\u003d19","jobTitle":"Data Analyst / Business Systems Analyst (Data Warehouse)","company":"Agile Global Solutions, Inc","location":"Denver,

CO","date":"2017-02-
25"},{"detailUrl":"http://www.dice.com/job/result/10441189/27592?src\u003d19","jobTitle":"Business Systems Analyst III - Claims","company":"Fidelis Care New York","location":"Buffalo, NY","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/10105937/696702?src\u003d19","jobTitle":"Business Systems Analyst","company":"DEEGIT, INC.","location":"Greenville, SC","date":"2017-02-22"},{"detailUrl":"http://www.dice.com/job/result/10124587/2203?src\u003d19","jobTitle":"Senior Systems Analyst","company":"xScion Solutions","location":"Washington, DC","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/10113579/580597?src\u003d19","jobTitle":"Systems Analyst","company":"Matlen Silver Group","location":"Charlotte, NC","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/cxelisen/260944?src\u003d19","jobTitle":"Senior Systems Analyst","company":"Eliassen Group","location":"Dulles, VA","date":"2017-02-24"}]}

About the Authors

Vlad Krotov is an Associate Professor and MSIS Program Director at the Department of Computer Science and Information Systems, Arthur J. Bauernfeind College of Business, Murray State University. He received his PhD in Management Information Systems from the Department of Decision and Information Sciences, University of Houston (USA). Prior to joining Murray State University, he served as an Associate Professor and Assistant Dean of Graduate Programs at Abu Dhabi University (UAE). His teaching, research, and consulting work is devoted to helping managers use data and technology for creating organizational value. His research has appeared in a number of academic and practitioner-oriented journals: *CIO Magazine*, *Journal of Theoretical and Applied E-Commerce*, *Communications of the Association of Information Systems*, *Business Horizons*, *Journal of Competitiveness Studies*, *Polish Journal of Management Studies*, *Journal of Cases in Educational Leadership*, *Journal of Emerging Technologies in Accounting*, *The Electronic Journal of Information Systems Evaluation*, *Journal of Education for Business*, *Polish Journal of Management Studies*, etc. His research was recognized at Murray State University by the “Outstanding Research Award” in 2016 the "Emerging Scholar Award” in 2017.

Leigh Redd Johnson is a Professor of Business Ethics and Law in the Accounting Department of Murray State University where she has taught since 2006. Prior to joining Murray State, Ms. Johnson was a corporate and securities associate at Womble Carlyle Sandridge & Rice, PLLC in its Research Triangle Park, North Carolina office. She has published in *Issues in Accounting Education, Journal of Accounting Education, Research on Professional Responsibility and Ethics in Accounting, The U.C. Davis Business Law Journal, The Journal of Corporate Accounting and Finance, The Journal of Corporate Taxation, The American Journal of Business Education, and Journal of Legal, Ethical and Regulatory Issues*, among others.

Leiser Silva has served as a faculty member at the Bauer College of Business at the University of Houston since 2002 and serves as the Associate Dean for Bauer's Graduate and Professional Programs. Silva holds a Ph.D. in Information Systems from the London School of Economics and Political Science, a M.Sc. in Systems Analysis and Design from the same institution, and a B.Sc. in Computer Sciences from the Universidad del Valle in Guatemala. His current research examines issues of power and politics in the adoption and implementation of information systems, particularly in the context of public organizations. Silva has made contributions to the body of literature in Information Systems that studies the relationship between strategy and Information Technology. His work has been published in leading academic journals such as *MIS Quarterly*, *Journal of the Association of Information Systems (JAIS)*, *Journal of Information Technology (JIT)*, *European Journal of Information Systems* and *Information Systems Journal*.