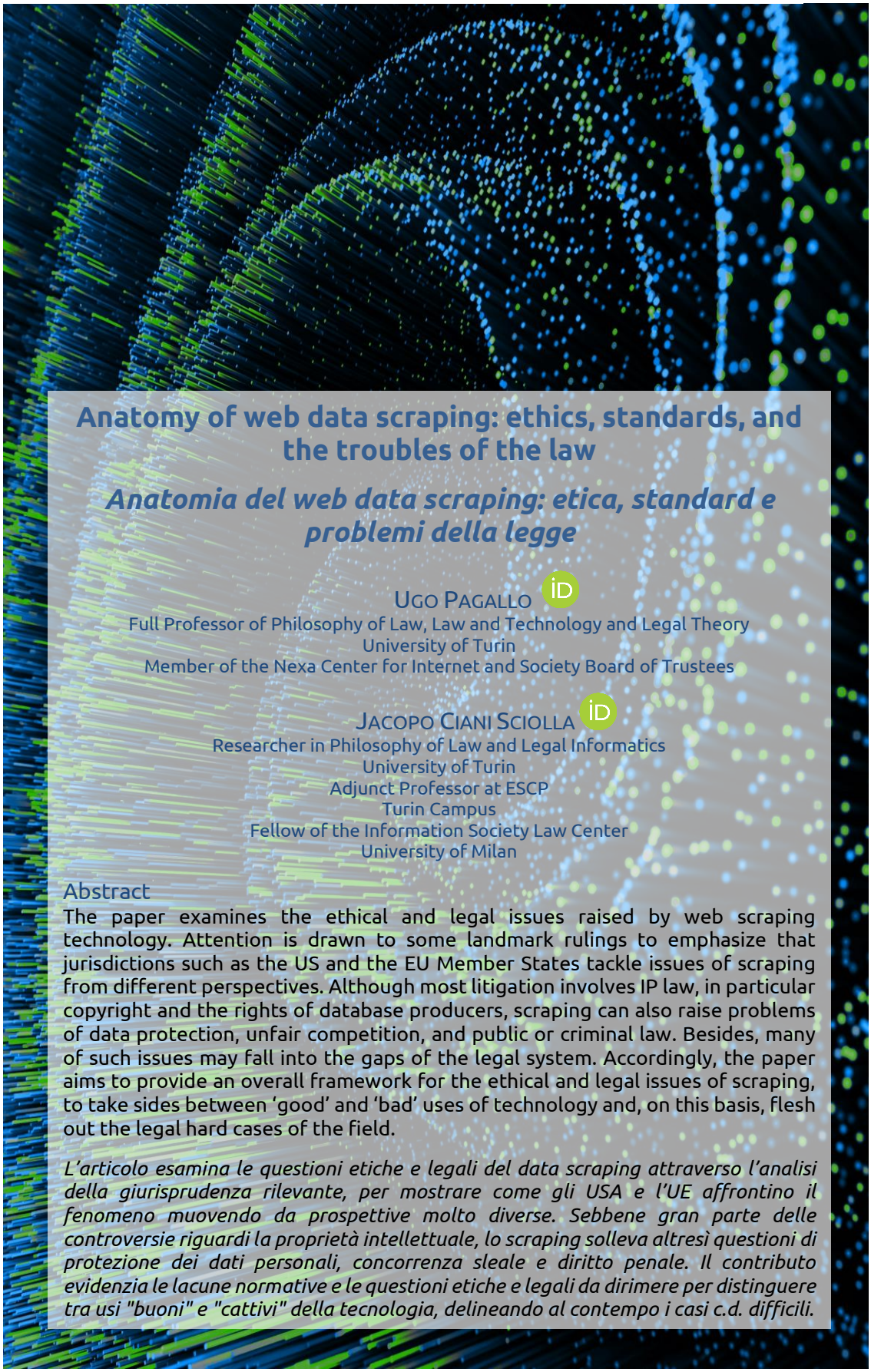
Pagallo U, Ciani Sciolla J, 'Anatomy of web data scraping: ethics, standards, and the troubles of the law' (2023) 2 EJPLT, 1 - 19.
DOI: https://doi.org/10.57230/EJPLT232PS

EJPLT
ISSN 2704-8012



Anatomy of web data scraping: ethics, standards, and the troubles of the law

Anatomia del web data scraping: etica, standard e problemi della legge

UGO PAGALLO
Full Professor of Philosophy of Law, Law and Technology and Legal Theory
University of Turin
Member of the Nexa Center for Internet and Society Board of Trustees

JACOPO CIANI SCIOLLA
Researcher in Philosophy of Law and Legal Informatics
University of Turin
Adjunct Professor at ESCP
Turin Campus
Fellow of the Information Society Law Center
University of Milan

Abstract

The paper examines the ethical and legal issues raised by web scraping technology. Attention is drawn to some landmark rulings to emphasize that jurisdictions such as the US and the EU Member States tackle issues of scraping from different perspectives. Although most litigation involves IP law, in particular copyright and the rights of database producers, scraping can also raise problems of data protection, unfair competition, and public or criminal law. Besides, many of such issues may fall into the gaps of the legal system. Accordingly, the paper aims to provide an overall framework for the ethical and legal issues of scraping, to take sides between 'good' and 'bad' uses of technology and, on this basis, flesh out the legal hard cases of the field.

L'articolo esamina le questioni etiche e legali del data scraping attraverso l'analisi della giurisprudenza rilevante, per mostrare come gli USA e l'UE affrontino il fenomeno muovendo da prospettive molto diverse. Sebbene gran parte delle controversie riguardi la proprietà intellettuale, lo scraping solleva altresì questioni di protezione dei dati personali, concorrenza sleale e diritto penale. Il contributo evidenzia le lacune normative e le questioni etiche e legali da dirimere per distinguere tra usi "buoni" e "cattivi" della tecnologia, delineando al contempo i casi c.d. difficili.

Keywords: data ownership; data scraping; web harvesting; data grabbing; intellectual property; data protection; trade secret; data access; data ethics; standards.

Introduction to data extraction via web scraping.

*The notion of 'data scraping' can be understood as the extraction of data from a website to process, analyze, and present such data as useful information, especially for commercial purposes. Web services gather information from data hosts-websites that store or house target data, either by parsing or by scraping such data. Parsing generally refers to the collection of information through a series of formalized data requests and using application programming interfaces ('APIs'). Data structures, like Extensible Markup Language, or XML feeds, suited for automated computer processing, provide the format through which data can be efficiently stored, searched, and shared. However, such data interchange may not be available because of opposing interests of data holders or owners. Data scraping circumvents this problem by automatically extracting useful information from the Hyper Text Markup Language ('HTML') code that most websites display. This automated process, also called web crawling, harvesting, or data grabbing, is carried out by bots, web crawlers or web spiders.

Web scrapers typically take something out of a webpage to use it for another purpose, e.g., copying names and phone numbers, or companies and their URLs, to make a list, known as 'contact scraping'. Popular metasearch engines or search aggregators, such as Skyscanner and Booking.com, hinge on data scraping software to retrieve and aggregate search results of online travel agencies. Since these services depend on third parties' data, data holders often deem screen scraping illegitimate. As a result, a battle between website developers and scraping developers has followed, and still, from a legal viewpoint, the terms of this battle look uncertain.

* Ugo Pagallo wrote paras. 2 and 4; Jacopo Ciani Sciolla wrote paras. 3.1-3.6. All authors wrote paras. 1 and 5 and have read and agreed to the published version of the manuscript.

Correspondingly, the paper aims to deepen the normative issues[1], i.e., both moral and legal, brought about by data scraping, according to a twofold approach. On the one hand, the focus is on applied ethics and both the good and bad uses of technology. Data scraping does not only concern possible infringements of rights, e.g., the rights of data owners and website developers, but also several uses of technology that are arguably beneficial. On the other hand, attention is drawn to the current state-of-the-legal-art. In particular, the focus shall be both on the US and Europe which, while simplifying a more complex global scenario, provide a quite comprehensive overview of the array of legal protections against, e.g., screen scraping and allow to make our findings and recommendations equally potentially relevant to other countries. Landmark rulings across multiple jurisdictions illustrate the key legal claims brought against data scrapers and how such claims, however, often fall within the loopholes of the law. Drawing on this basis, in the final part of this paper, a list of open issues that will need further consideration is under scrutiny, namely the pros and cons of the technology, the rulings of the Courts and the persisting legal uncertainty. Despite different approaches and opinions among scholars and jurisdictions, the aim is to cast light on some common trends in the legal domain that offer a guide to the normative troubles of the field.

1. The ethics of data scraping.

Scraping has many useful applications, and it is often used by individuals serving the public interest. Many widespread uses of data scraping benefit both data hosts and scrapers. Scraping services allow users to find the information they seek more easily. Journalists use scraping technology to gather and analyze massive chunks of statistical data. Scholars employ scraping technology for their academic research. Popular search engines, such as Google and Bing, crawl the web and scrape web pages to provide search results. Since most website traffic often comes from search engines, data hosts do not oppose access to most crawling bots, given the overwhelming benefit they derive. In the case of metasearch engines, data scraping enables them to extend the gathering of data and information – and to compare e.g., product and price information from various sources – exempting users from visiting and checking multiple webpages or search engines. Further possible uses of the technology include tracking companies' reputation or aggregating news and other content on curated websites.[2] However, despite these beneficial applications, scraping technology can also be used for malicious purposes, such as spamming email accounts, causing website crashes,[3] or setting scams.[4]

[1] An overview of legal issues arising from data scraping is provided also in J Ciani Sciolla, 'The normative challenges of data scraping: legal hurdles and steps forward' (2023) 16 i-lex Riv Scienze Giuridiche, Scienze Cognitive ed Intelligenza Artificiale: i-vi, working as an introduction to a special issue entirely devoted to study the many legal facets of the phenomenon.

[2] A Sellars, 'Twenty Years of Web Scraping and the Computer Fraud and Abuse Act' (2018) 24 BU J Sci & L Tech, 381-382.

[3] M Boulanger, 'Scraping the Bottom of the Barrel: Why It Is No Surprise That Data Scrapers Can Have Access to Public Profiles on Linkedln' (2018) 21 Smu Sci & Tech L Rev,77-78.

[4] K Collier, 'Why Cybercriminals Looking to Steal Personal Info Are Using Text Messages as Bait'. NBC NEWS, 6 May 2021.

Scraping may also be parasitic: scrapers can benefit from the exclusion or detriment of data hosts; they can undercut a website's revenue by republishing scraped data without requiring users to view supporting advertisements; moreover, scrapers can attain their own ad revenues, viewers, and customers by taking content from another data host. Scraping may also affect the protection of fundamental rights by collecting personally identifying information with serious privacy implications. For example, the technology is problematically employed in the so-called 'mugshot gallery': by displaying the photos of arrested people, scrapers monetize them in various ways, e.g., charging individuals to have their images removed.[5]

Against this backdrop, scholars often insist on both 'beneficial' and 'harmful' uses of scraping,[6] also referring to data scraping as a phenomenon 'intertwined with both good and evil' and with 'paradoxical two-sided effects'.[7] Arguably, the troubles with scraping do not regard the technology as such. Data scraping is a form of copying that simulates the human processing of copying and pasting texts or images from a webpage. In general terms, scholars have extensively discussed the 'ethics of copying'.[8] Copying is essential for individual and social learning processes, cultural development, and economic success.[9] Copying enables democratization processes by providing access to cultural goods and relevant information. Mental activity seems to be a form of copying as well. Therefore, it is simply wrong to assume that copying should be regarded as something immoral.[10] Rather, attention should be drawn to how copying fits into a broader system of values that scholars have explored in such fields as philosophy of art and technology, philosophy of law and ethics, legal theory and media studies, art history and literary theory, or sociology.[11] Drawing on this basis, it is thus evident that the problem does not regard any scraping as such. The assumption is confirmed by the paradoxical effects that would follow a ban on technology. The overall prohibition of web scraping would increase the barriers-to-entry by prescribing licensing deals for new entrants. In addition, every ban could negatively impact competition and hence foster monopolistic positions of incumbents.[12] Correspondingly, what is at stake is not the 'morality of scraping' but rather how to strike a fair balance between such web practices and a broader system of values, such as the rights and interests protected under legal provisions[13]. After all, the activity of

[5] EK Lee, 'Monetizing Shame: Mugshots, Privacy, and the Right to Access' (2018) 70 Rutgers U L Rev: 566, 569; A Rostron, 'The Mugshot Industry: Freedom of Speech, Rights of Publicity, and the Controversy Sparked by an Unusual New Type of Business' 90 Wash U L Rev,1323-1324.
[6] MF Din, 'Breaching and Entering: When Data Scraping Should Be a Federal Computer Hacking Crime' (2015) 81 Brooklyn L Rev, 412-415.
[7] L Qian, J Tao, 'Rethinking Criminal Sanctions on Data Scraping in China Based on a Case Study of Illegally Obtaining Specific Data by Crawlers' (2020) 4 China Leg Science,136-158.
[8] R De George, 'Information technology, globalization and ethics' (2006) 8 Ethics and Inf Tech:29, 40; DH Hick, R Schmücker (Eds) *The Aesthetics and Ethics of Copying* (Bloomsbury Publishing, 2016).
[9] U Pagallo, 'The Troubles with Digital Copies: A Short KM Phenomenology' In *Digital Rights Management: Concepts, Methodologies, Tools, and Applications* (Hershey, IGI Global 2013) 1379, 1394.
[10] S Bringsjord, 'In Defense of copying' (1989) 3 Public Aff Quart, 1-9.
[11] Hick and Schmücker (n 8).
[12] M Husovec, 'The end of (meta) search engines in Europe?' (2014) *Max Planck Inst for Innov and Comp Res Pap Series*, 14-15.
[13] V Krotov, L Johnson and L Silva, 'Legality and Ethics of Web Scraping' (2020) Communications of the Association for Information Systems, 47.

copying in itself acquires different moral connotations according to the context. Therefore, an ethical approach is necessary to discern good from bad uses of technology, especially considering the troubles of the law with the very distinction between legitimate and illegitimate web scraping practices. The focus of the following parts of this paper is on such current legal troubles and how ethics is relevant for the assessment and further development of current legal standards.

2. The law of data scraping.

The 'ethics of copying' shall be tested against the myriad legal problems brought about by data scraping. No statute specifically addresses the challenges of technology. Multiple jurisdictions – e.g., the US and EU national courts case law, under scrutiny in this paper – have followed a wide range of different paths. They regard legal issues on (i) unfair competition of metasearch engines; (ii) contractual liability as regards the website terms and conditions and whether they are binding for the scraper; (iii) non-contractual liability, with critical differences between common law and civil law jurisdictions; (iv) copyright; (v) sui generis rights on databases; down to, (vi) personal data, or data which is kept secret. Each of these cases poses thorny legal issues of its own. Let us start the legal analysis of this part of the paper with the lenses of tort law on unfair competition. The notion is traditionally intertwined with ethical considerations on whether a defendant may have violated 'honest practices in industrial or commercial matters' (art. 10 bis TRIPs agreement).

2.1. Unfair competition of metasearch engines, or the Ryanair saga.

In the introduction, we mentioned the battle of website developers against scraping developers, such as popular metasearch engines or search aggregators that hinge on third-parties' data. We noted that data owners, or data holders, often deem screen scraping as illegitimate. For example, Ryanair, the low-cost airline company headquartered in Ireland, lodged multiple parallel lawsuits against online travel agencies in the early 2010s. The aim was to stop the scraping of information from Ryanair's website that allowed such online travel agencies to sell flight tickets through their own services. In December 2013, the Court of Justice of the European Union handed down the *Innoweb* ruling. The Court labeled the actions of the metasearch engines as 'nearly parasitic', e.g., letting users parse through multiple databases of car ads listed on third-party sites (para. 48). In addition, Justices in Luxembourg held that metasearch engines likely infringe the database right of the indexed website, assuming that such right represents a protectable subject matter of its own.

It is noteworthy, however, that several national courts in Europe did not follow suit. For example, the German Federal Supreme Court, in *Ryanair v. Vtours*, established that a metasearch engine that carries out a booking on behalf of the consumer, or even under its own name (integrated booking), does not necessarily engage in an act of unfair competition. Likewise, in *Ryanair v.*

Viaggiare,[14] the Italian Supreme Court stated that scraping datasets is not illegal per se, except for possible infringements of intellectual property rights (IPRs).[15] The same holds true for the Grand Instance Court in Paris and the Spanish Court of Appeal, which established that travel agencies did not incur in the tort of parasitic unfair competition since their operations did not affect the normal functioning of the market or alter the market's competitive structure. Furthermore, the business of such online companies even brought Ryanair new customers and was overall beneficial for both the users and the market. The opposite views of the EU Court of Justice and several national jurisdictions in Europe are not unique to the old continent. Further cases on web scraping in different US Circuits have, in fact, seen opposite conclusions on how to interpret the Computer Fraud and Abuse Act ('CFAA') regarding data scraping practices. The following subsections examine this disagreement in detail, distinguishing between contractual and non-contractual liability of scraping developers and between civil and common law jurisdictions.

2.2. Contractual liability: Are the website terms and conditions binding for the scraper?

Data holders' claims against web scrapers involve both contractual obligations and non-contractual duties. Under the first scenario, a website's terms of service, or 'ToS' may prohibit scraping[16]. This was the case with the Ryanair's ToS. They allowed screen-scraping only for non-commercial purposes, thus excluding flight ticket sales in competition with Ryanair. Since there is no statute – as far as we know – that specifically targets data scraping, plaintiffs have intended to challenge the behavior of scraping developers on contractual basis. Since most websites' ToS prohibit the automated collection, scraping, use, and reproduction of data without permission, the question is how courts may interpret these terms, in particular, whether such ToS should be conceived of as enforceable contracts[17].

Courts and scholars usually distinguish three types of online contracts: (i) clickwrap, i.e., the user must affirmatively click 'I agree' before accessing the website; (ii) scrollwrap, i.e., the user must scroll through the contract and click 'I agree'; and, (iii) browsewrap, i.e., the user agrees to the contract by using the website.[18] Courts generally find clickwrap and scrollwrap agreements enforceable because, at some point during the registration or use of the website, users check a box to acknowledge they agree to the terms. However, since, most of the time, scraping activities do not require any consent to ToS – apart from browsewrap contracts or similar – the enforceability of ToS is problematic. All in all, both in Europe and in the US, courts have dismissed breach of contract claims against web scrapers when there is no evidence that

[14] Cass Civ 18.12.2018 n 2289, 2290.
[15] FE Beneke Avila, 'The German Federal Supreme Court's judgment in Booking.com as a case study of the limitations of competition law' (2022) IIC Int Rev IP and Comp L,1374.
[16] A Quarta, MW Monterossi, 'Web Scraping: A Private Law Perspective' (2023) 16 i-lex Riv Scienze Giuridiche, Scienze Cognitive ed Intelligenza Artificiale, 46-52.
[17] M Borghi, S Karapapa, 'Contractual Restrictions on Lawful Use of Information: Sole-Source Databases Protected bythe Back Door? ' (2015) 37 EIPR, 505-511.
[18] E Canino, 'The Electronic "Sign-in-Wrap" Contract' (2016) 50 U.C. Davis L. Rev, 539-541.

the link to the terms of use is displayed among all the links of the site, at least, at the bottom of the page as most sites on the internet.[19] Moreover, web scrapers are not accountable when they have no actual or 'constructive knowledge' about the terms and conditions of the website.[20] The Grand Instance Court in Paris found that Opodo was not bound by the terms of service on the Ryanair website; as much as the Spanish Supreme Court concluded that accessing and visiting the Ryanair website – free to anyone who types the URL address – does not entail any consent to enter into a contract. The Italian Supreme Court ruled along the same lines, pointing out that, although the scraper carried out something not allowed by Ryanair's ToS, the scraper never consented to such ToS, and no contract violation would thus exist. However, we should not jump to conclusions. Other cases show that plaintiffs can demonstrate the 'actual knowledge' of their counterparties through cease-and-desist letters. For example, in *Ryanair Ltd v Billigfluege.de GmbH*, Ireland's High Court ruled that Ryanair's 'click-wrap' agreement is legally binding since the hyperlink was plainly visible, putting the burden of agreeing to the ToS in order to gain access to the online services on the user. The ruling of the Irish Court over the Irish company – and against a German travel agency – seems at odds with most jurisdictions in the old Continent. This legal uncertainty has suggested that data owners, or data holders, such as Ryanair, complement their contractual claims with further claims of non-contractual liability.

2.3. Non-contractual liability: Common law versus civil law approaches toward data scraping.

Non-contractual liability issues of data scraping highlight the differences between common law and civil law. In Europe, most of the time, courts have to decide whether the scraper infringed exclusivity rights over the scraped data: intellectual property rights are the natural legal framework for such cases in EU law. Therefore, the focus of EU lawyers is on whether scraping extracts data that is protected as a database. Other laws, e.g., data protection law, competition law or the law governing the use of public sector information, may also be relevant, depending on the features of the scraped data. In the US, the 'thin' protection provided to databases by the so-called 'copyright's database-sized gap' has recommended website owners seek remedies in criminal and tort law. In particular, discussions have often revolved around the CFAA from 1986, which punishes 'whoever ... intentionally accesses a computer without authorization or exceeds authorized access, and thereby obtains ... information'.

---

[19] In *QVC Inc. v. Resultly LLC*, No. 14-06714 (E.D. Pa., filed 24 November 2014) the court ruled that the ToS should be brought to the users' attention in order for a browse wrap contract or license to be enforced. This is not the case when the ToS link is displayed among all the links of the site, at the bottom of the page as most sites on the internet.

[20] For example, in *Cvent, Inc. v. Eventbrite, Inc.*, 739 F. Supp. 2D 927, 96 U.S.P.Q. 2d 1798 9 (E.D. Va. 2010), a federal court in the Eastern District of Virginia dismissed a breach of contract claim against a web scraper because the plaintiff's T&C was a browsewrap agreement and the plaintiff had not 'pled sufficient facts to plausibly establish that defendants Eventbrite and Foley were on actual or constructive notice of the terms and conditions posted on Cvent's website.'; the court reached a similar conclusion in *Alan Ross Machinery Corporation v. Machinio Corporation*, No. 1:2017cv03569 - Document 31 (N.D. Ill. 2018).

For years, courts have debated the meaning of the formula 'without authorization' under the CFAA, i.e., whether the CFAA is best understood as 'an anti-intrusion statute' or a 'misappropriation statute'.[21] In 2019, the Ninth Circuit issued its opinion in *hiQ Labs, Inc. v. LinkedIn Corporation,*[22] in which the key question was to determine whether a computer's gates were up or down. The Court interpreted the CFAA as an anti-intrusion statute prohibiting unauthorized access and what is traditionally conceived as 'hacking'.[23] Unauthorized access occurs 'when a person circumvents a computer's generally applicable rules regarding access permissions, such as username and password requirements'. Public websites like LinkedIn, however, permit public access to their data, such as available LinkedIn member profiles, so that the decision of the Ninth Circuit was that 'a user's accessing that publicly available data will not constitute access without authorization under the CFAA'. In light of this ruling, not only have several scholars argued that the CFAA is ill-suited for tackling web scraping cases,[24] but also, far from relying on the protection of the CFAA rules, plaintiffs have embraced further legal strategies under US tort law, so that web scraping practices would entail either common law misappropriation, i.e., the 'hot news' tort;[25] unjust enrichment;[26] conversion;[27] or trespass to chattel.[28] The 'hot news' tort originated with the seminal Supreme Court case *INS v. AP*, in which a press agency copied the facts reported in the news bulletins of its competitor. Despite facts in the news being considered uncopyrightable information in the public domain, the Supreme Court recognized that these facts were 'the result of organization and the expenditure of labor, skill, and money'. Correspondingly, *INS* was 'endeavoring to reap where it has not sown'.[29] The 'hot news' tort thus applies to cases of web scraping and protected websites according to four conditions: (i) the plaintiff generates or collects information at some cost or expense; (ii) the defendant's use of the information constitutes free-riding on the plaintiff's costly efforts to generate or collect it; (iii) the defendant's use of the information is in direct competition with a product or service offered by the plaintiff; and, (iv) the ability of other parties to free-ride on the efforts of the plaintiff would shrink the incentive to create the product or service, whose existence or quality would then substantially be threatened.[30]

The problem with the 'hot news' tort approach, however, is that it only protects time-sensitive information, leaving a wide variety of databases unprotected from scraping. To prevent this lack of protection, two further state law torts, i.e., conversion and trespass to chattel, may provide some

[21] M Addicks, 'Van Buren v. United States: The Supreme Court's Ruling on the Fate of Web Scraping - "Access" to Discovery or Detention?' (2022) 24 Tul J Tech & Intell Prop, 161.
[22] 938 F.3d 985, 995 (9th Cir. 2019).
[23] Boulanger (n 3).
[24] JE Christensen, 'The Demise of the CFAA in Data Scraping Cases' (2020) 34 Notre Dame J.L. Ethics & Pub. Pol'y, 529.
[25] See e.g. *Allure Jewelers, Inc. v. Ulu*, No. 1:12CV91, 2012 WL 4322519 (S.D. Ohio 25 September 2012).
[26] See e.g. *Snap-on Bus. Sols. Inc. v. O'Neil & Assocs, Inc.*, 708 F. Supp. 2D 669, 680-82 (N.D. Ohio 2010); *ShopLocal LLC v. Cairo, Inc.*, No. CIV.A. 05 C 6662, 2006 WL 495942, at *3 (N.D. Ill. 27 February 2006).
[27] See e.g. *QVC, Inc. v. Resultly,* LLC, 159 F. Supp. 3D 576(E.D. Pa. 2016).
[28] See e.g. eBay, *Inc. v. Bidder's Edge, Inc.*, 100 F. Supp. 2d 1058, 1065-71 (N.D. Cal. 2000).
[29] *Int'l News Serv. v. Associated Press*, 248 U.S. 215, 247 (1918).
[30] *Nat'l Basketball Ass'n v. Motorola, Inc.*, 105 F.3d 841, 852-53 (2d Cir. 1997).

remedy. Liability for trespass to chattels traditionally concerns the intention to take or intermeddle with a chattel possessed by someone else. Since the 1990s, this tort has been applied to cases involving devices that automatically overuse phone and email networks and diminish their functionality. The tort has been used in scraping cases as well. To establish a trespass to chattels claim, data hosts have to show that they were dispossessed of their chattel, that their chattel's condition, quality, or value was impaired, or that they were 'deprived of the use of their chattel for a substantial time.' In the case of conversion, the difference with the trespass to chattels claim consists in a more serious deprivation of the owner's rights, so that 'an award of the full value of the property is appropriate.'

These torts and the CFAA can be understood as 'quasi-IP' regimes because they serve to protect proprietary interests which are not covered by the law of copyright. To demonstrate conversion or trespass to chattel, plaintiffs must meet a stringent standard regarding the proof of actual injury to their computer systems. In many cases, however, scraping will have a negligible impact on the plaintiff's servers, making these claims unlikely to succeed. By considering the drawbacks of US regulations, would the EU sui generis right on databases offer better protection?

2.4. Copyright over data and datasets.

The question of whether and to what extent databases should be protected by the law concerns two potentially conflicting goals: that of providing adequate incentives for the continued production of such databases, and that of ensuring public access to the information they contain. In Europe, the assumption is that the disparity of the fixed costs needed to create a database and the marginal costs to copy or access it, necessarily makes it an *ad hoc* intellectual property right, i.e., Directive 96/9 on the legal protection of databases. The protection entails an exclusive right that enables the database maker to charge, for a limited time, a price superior to the marginal cost for the use of its database and to select other market participants who can take advantage of such database as licensees, thus boosting the database production.

Things have gone differently in the US.[31] Based on the 1998 US Copyright Office's report on legal protection for databases, Congress dismissed any proposals for establishing a new form of database right on the grounds that copyright law should not favor the protection of factual data unless facts have been selected, coordinated, or arranged in an original way. Therefore, the threshold is not satisfied when the setting up of a database is determined by technical considerations, rules or constraints, which leave no room for creative freedom.[32] The US Supreme Court has expressly rejected the 'sweat of the brow' theory, under which copyright could be granted just as 'a reward for the hard work that went into compiling facts'. In its seminal case, *FeistPublications*

[31] JC Ginsburg, 'Copyright, Common Law, and Sui Generis Protection of Databases in the United States and Abroad' (1997) 66 U Cin L Rev, 153-157.
[32] Case C-604/10, *Football Dataco Ltd vs. Yahoo! UK Ltd* (CJ, 1 March 2012).

Inc. v. Rural Telephone Service Co.,[33] Justices in Washington denied copyright protection to a telephone directory because the latter was a mere collection of unoriginal facts, regardless of whether immense efforts were needed to compile the directory. The US District Court of the Central District of California confirmed this view in *Ticketmaster v. Tickets.com,* [34] where Tickets.com employed an electronic web crawler that reviewed Ticketmaster's website and extracted the information on events for which Ticketmaster offered the sale of tickets. The court reasoned that 'the primary star in the copyright sky... is that purely factual information may not be copyrighted.'

In addition, although data scraping may amount to *prima facie* copyright infringement, it could still fall under copyright exceptions, limitations, or fair use defenses. In *Perfect10 v. Amazon.com,*[35] for example, Google was sued under copyright law for scraping thumbnail images of a magazine's covers as a part of the Google Image Search. However, the Ninth Circuit found Google's scraping to be fair use because it 'provided a significant benefit to the public' and was highly transformative.

In Europe, where copyright law does not include a fair use clause, scraping may be lawful under a series of 'exceptions.' They regard (i) the exception for temporary copies (Art. 5.1 Directive 2001/29); (ii) the text and data mining exception introduced by Art. 3 and 4 of the Directive on Copyright in the Digital Single Market (CDSM);[36] down to Art. 1.6 of Directive (EU) 2019/1024 on open data and the re-use of public sector information. According to this latter directive, 'the right for the maker of a database provided for in Article 7(1) of Directive 96/9/EC shall not be exercised by public sector bodies in order to prevent the re-use of documents or to restrict re-use beyond the limits set by this Directive'. Leaving aside the interpretation of these provisions, it is noteworthy that some scholars are pessimistic about the possibility of anyone who is not a research and cultural organization acting for research purposes lawfully scraping data in Europe.[37] Yet, in Italy, the Court of Rome has considered the scraping of the state rail-transport operator's data conducted by a web aggregator lawful on the basis of Art. 7(1) of the open data directive.[38] In light of these 'exceptions', that protect web scrapers activities, the focus is next fixed on the corresponding general rules.

---

[33] *Feist Publ'ns, Inc v. Rural Tel Serv Co.*, 499 U.S. 340, 345 at 361 (1991).

[34] *Ticketmaster Corp. v. Tickets.com, Inc*, 2003 WL 21406289 (CD Cal. 2003). See S O'Reilly, 'Nominative Fair Use and Internet Aggregators: Copyright and Trademark Challenges Posed by Bots, Web Crawlers and Screen Scraping Technologies' (2007) 19 Loy Consumer L Rev, 273.

[35] 508 F.3d at 1155-56 (9th Cir. 2007).

[36] The goal of Art. 3 is to introduce a mandatory exception under EU copyright law which exempts acts of reproduction (for copyright subject matter) and extraction (for the Sui Generis Database Right) made by research organisations and cultural heritage institutions in order to carry out text and data mining for the purposes of scientific research. Art. 4 mirrors Art. 3 with one major difference: it is available to any type of beneficiaries for any type of use but can be expressly reserved by rightsholders – in other words it may be the object of 'opt-out' or 'contract-out'. C Gallese, 'Web scraping and Generative Models training in the Directive 790/19' (2023) 16 i-lex Riv Scienze Giuridiche, Scienze Cognitive ed Intelligenza Artificiale, 1-13.

[37] T Margoni, M Kretschmer, 'A deeper look into the EU Text and Data Mining exceptions: Harmonisation, data ownership, and the future of technology' (2021) CREATe Working Paper, 7.

[38] J Ciani Sciolla, *Il pubblico dominio nella società della conoscenza. L'interesse generale al libero utilizzo del capitale intellettuale comune* (Giappichelli 2021), 221.

2.5. The database sui-generis right.

In EU law, the assumption is that databases shall not be original to be protected. Contrary to US law, Chapter III of Directive 96/9/EC grants a 'sui generis' protection to EU makers of a database, which shows that there has been a qualitatively and/or quantitatively substantial investment in either the obtaining, the verification or the presentation of the contents. In particular, the database creator is entitled to prevent extraction and/or re-utilization of both the whole database or a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database (Art. 7). If the requisites are met, the protection is granted automatically for 15 years starting either from the creation date or from the day in which the database was first made publicly available.

Since the early days of the database sui generis rights, several plaintiffs have progressively invoked this law against the data scraping of metasearch engine operators. Consequently, courts must first check whether the scraped website contains a database according to the definition of Art. 1.2 of the database directive,[39] whereas Art. 3.2 clarifies that the protection does not extend to the contents of the database. On this basis, it is worth mentioning that in one of the multiple lawsuits brought forth by Ryanair in Europe, the Spanish Supreme Court ruled that Ryanair does not have a collection of independent data under the database provisions, but rather, rights upon a software that generates the information requested under the parameters introduced by the user, i.e., a software that provides the best price for the flights users are looking for, considering a range of variable factors.

The ruling of the Spanish Supreme Court, however, has been contradicted by other courts. For example, the Italian Supreme Court[40] and the Regional Court of Hamburg[41] have both declared that Ryanair does have a database under Art. 1.2 of the 1996 Directive. The assumption begs two further questions concerning the main legal thresholds that trigger the protection of a database: (i) substantiality of the investment; and (ii) substantiality of the extraction.

2.5.1. Substantiality of the investment.

The first threshold is defined as the 'qualitatively and/or quantitatively substantial investment in either the obtaining, verification or presentation of the contents' (Art. 7(1) of the database directive). According to the CJEU, what matters is that 'the obtaining of those materials, their verification or their presentation [...] required substantial investment in quantitative or qualitative terms, which was independent of the resources used to create those materials.' The burden to prove all the facts substantiating the relevant investment falls upon the person or company claiming protection.

---

[39] 'For the purposes of this Directive, "database" shall mean a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means.'

[40] Above n 14.

[41] *Ryanair v Cheaptickets* (26 February 2010).

EU national courts have interpreted this requirement in different ways. For example, the Grand Instance Court in Paris found no infringement of a database right as Ryanair did not prove to have made substantial investments necessary to pretend such protection. Likewise, as mentioned in the previous section, the Spanish Supreme Court concluded that Ryanair's substantial investment was not directed to protect data but to create a software that generates information under the user's search query parameters.[42] Similar arguments and outcomes are stressed in *Ryanair v. eDreams,*[43] another legal case that ended with a dismissal of Ryanair's claims. On the contrary, the Italian Supreme Court considered that Ryanair's investment can be presumed, although this presumption does not entail that the claims of the Irish company should be deemed substantiated. Against these opposite views, attention should be drawn to the second threshold of the EU database directive.

2.5.2. Substantiality of the extraction.

The second threshold requires that the acts to extract or reuse data via scraping concern the totality or a substantial part of the database. Extracting or reusing data in a limited or partial way – i.e., in both quantitative and qualitative terms – is thus implicitly considered admissible, regardless of any rightsholder's authorization. Also, scraping is lawful when those acts are systematically and repeatedly carried out but involve non-substantial parts of the information contained in the database, i.e., the so-called 'diachronic extraction.' Vice versa, no extraction or reuse is allowed insofar as such acts require operations that conflict with the normal use of the database or cause excessive harm to the legitimate interests of its creator(s). Accordingly, courts are requested to establish whether the extraction is 'substantial enough' or – in the event of diachronic extraction – the latter harms the legitimate interest of the database maker.

Interestingly, these requirements are also taken into account by US Courts, although for different purposes. On the basis of the substantial copying requirement, courts have to determine whether competition is fair or unfair, that is, whether scraping triggers beneficial dissemination of information or harmful free-riding. The Eleventh Circuit, for instance, held that scraping is unlawful only when it concerns a substantial amount of data, such that 'the block of data that the defendants took was large enough to constitute appropriation of the [d]atabase itself'.[44]

In order to assess the substantiality of the extraction and the related harm, some jurisdictions, e.g., Italy, inspect the logs of the scraped website. In a case where the number of logs corresponded roughly to 30% of the total daily accesses—a figure defined as 'non impressive'—the Court of Rome established that there was no extraction of a substantial part of the database.[45] However, the Court pointed out that such logs were legitimate only if they represented

[42] Tribunal Supremo, *Ryanair v Atrapalo*, Case No. 572/2012 (2012).
[43] Curtea de Apel Barcelona, R*yanair v Vacaciones Edreams* (2009).
[44] See e.g. *Compulife Software Inc v. Newman*, 959 F.3d 1288, 1314 (11th Cir. 2020).
[45] District Court of Rome, *Trenitalia S.p.a. v GoBright Media Ltd* (5 September 2019).

forms of contingent accesses and periodic and selective acquisitions of data, each of which following a specific search query by a user.

In the case of diachronic extraction, Italian courts have made it clear that the harm to the legitimate interest of the database maker should consist of something more than just the loss of profit or income that derives from the absence of a license and subsequent lack of payment of license fees. Since metasearch engines are considered pro-competitive businesses, the harm cannot result in any loss of revenues deriving from the reduction in the volume of the traffic on the scraped website. Instead, harm should be deemed unjust only when scraping impairs the database's normal operativity because of the overload caused by the scraper's traffic. This was the argument at stake in *QVC Inc. v. Resultly*, where the e-commerce site QVC claimed that the shopping aggregator 'excessively crawled' the QVC's retail site – allegedly sending 200-300 search requests to the QVC's website per minute, sometimes up to 36,000 requests per minute – so that the website crashed for two days, resulting in lost sales for QVC.

Most of the time, however, scraped companies do not lose anything as a result of these activities and, therefore, no proof of tangible harm can be demonstrated. In general terms, diachronic data extraction is far more difficult to challenge than synchronic extraction, which is not necessarily good news for data holders. Indeed, most scraping activities are diachronic.

2.6. The nature of scraped data: personal, public or secret.

The law applicable to scraping operations may vary depending on the nature of the scraped data. Such data can be personally identifiable information, publicly available information, or data and information kept under secret, e.g., trade secrets. Each case presents its specificities, which are examined separately in the following subsections.

2.6.1. The scraping of personal data.

Over the past few years, consumer data privacy legislation has gained traction in the US. The California Consumer Privacy Act (CCPA) and the California Privacy Rights Act (CPRA), for instance, regulate the collection of consumers' personal data and the sharing of such data with third parties.[46] However, it seems fair to admit that no currently proposed or enacted privacy statute adequately protects publicly available personal information. All of this is exempted, making it fair game to scrape, use, share, or sell such personal data. Therefore, US privacy laws have not been the frequent subject of web scraping litigation. Few scholars have indeed addressed the privacy implications of scraping publicly available personal information.[47]

On the other hand, the EU has much more stringent laws on the scraping of personal data, and EU law even imposes fines on web scrapers who violate

[46] W Barfield, U Pagallo, *Advanced Introduction to Law and AI* (Elgar 2020).
[47] G Xiao, 'Bad Bots: Regulating the Scraping of Public Personal Information' (2021) 34 Harv J L & Tech:702; AM Parks, 'Unfair Collection: Reclaiming Control of Publicly Available Personal Information from Data Scrapers' (2022) 120 Mich L Rev, 913.

provisions of data protection. The legal troubles of *Clearview AI* illustrate this point[48]. The company has been embroiled in multiple lawsuits stemming from its conduct in scraping billions of facial images from the Internet and creating a biometric database that allows users to immediately identify a member of the public merely by uploading a person's image to the database.[49]

In April 2020, the French data protection authority (CNIL) published guidance on the lawful scraping of personal data for direct marketing purposes. Although contact information may be available on publicly accessible websites, the individuals who posted the information do not reasonably expect to have it scraped for 'prospecting'. Therefore, according to CNIL, such personal data cannot be re-used for marketing without the consent of the data subject obtained prior to any reuse. The CNIL clarifies that accepting the ToS – meaning that the individual accepts to receive marketing communications – is insufficient, as it is not specific to make web scraping of personal data lawful.

On 24 August 2023, twelve data protection authorities from around the world published a joint statement outlining the key privacy risks associated with data scraping taking place on social media as well as the steps that both websites and individuals should take to minimize such risks and meet regulatory expectations.[50]

In general terms, the opinion of courts and regulators at the EU level suggests that web scraping involving the collection of personal data is unlawful if it does not comply with the principles and provisions of the general data protection regulation, the GDPR.[51]

2.6.2. The scraping of data as an essential facility.

Antitrust issues are often at the core of web scraping litigation. For example, antitrust issues with data control were at the core of the Ryanair saga in Europe. The Italian Supreme Court, among others, found that Ryanair abused its dominant position in the downstream market for the provision of information for its own flights (sole source) by refusing access to an essential facility it owns (data about flight schedules and prices). Similar problems may arise in the US as well. A monopolist operating a public website could be liable under the 'refusal to deal' doctrine under Section 2 of the Sherman Act for restricting data access by means of exclusionary data scraping prohibitions.[52]

In *hiQ Labs, Inc v LinkedIn Corp*, the Ninth Circuit questioned the legitimacy of LinkedIn's data scraping prohibition, explaining that 'if companies like LinkedIn, whose servers hold vast amounts of public data, are permitted selectively to ban only potential competitors from accessing and using that

[48] F Lala, 'Data collection via web scraping: privacy and facial recognition after Clearview' (2023) 16 i-lex Riv Scienze Giuridiche, Scienze Cognitive ed Intelligenza Artificiale, 34-44.
[49] I Neroni Rezende, 'Facial recognition in police hands: Assessing the 'Clearview case' from a European perspective' (2020) 3 *New Jour of Eur Crim L*,375-389.
[50] Joint Statement on data scraping and the protection of privacy, 24 August 2023.
[51] U Pagallo, 'The Legal Challenges of Big Data: Putting Secondary Rules First in the Field of EU Data Protection' (2017) 3 Eur Data Prot L Rev, 36.
[52] 15 USC § 2 provides that '[e]very person who shall monopolize, or attempt to monopolize, or combine or conspire... to monopolize...shall be deemed guilty of a felony.' See I Drivas 'Liability for Data Scraping Prohibitions under the Refusal to Deal Doctrine: An Incremental Step toward More Robust Sherman Act Enforcement' (2019) 86 U. Chi. L. Rev., 1901.

otherwise public data, the result - complete exclusion of the original innovator in aggregating and analyzing the public information - may well be considered unfair competition'. On this basis, the court dismissed LinkedIn's proffered justifications that the prohibition served its users' privacy because the company shared its users' data with third-parties for commercial purposes. Likewise, LinkedIn could not persuasively claim the prohibition was justified as a measure against free riding because the company could not claim ownership of the data and users clearly intended to make their profiles publicly accessible. The Ninth Circuit dismissed LinkedIn's defense because giving companies 'free rein to decide... who can collect and use data... risks the possible creation of information monopolies that would disserve the public interest.'.

2.6.3. The scraping of data kept secret.

Web scraping may also concern data protected as a trade secret. In EU law, the definition and protection requirements are set up in Directive 2016/943/EU (Trade Secrets Directive) on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure. Such provisions are very similar to the Uniform Trade Secrets Act ('UTSA') in US law. Both statutes are modeled on Art. 30 of the TRIPS Agreement and require both the existence of a trade secret and its misappropriation.

As regards the existence of such trade secret, it hinges on three conditions: (i) a trade secret must not be generally known to or readily ascertainable by others; (ii) the owner must have taken reasonable measures to keep such information secret; and, (iii) the information must originate independent economic value from the fact of being secret. The critical point that follows as a result is whether the use of data scraping technologies to automatically extract the information contained in a website makes this information 'readily accessible' and, hence, not so secret. Information is considered readily accessible if it is not generally known but can be obtained by otherwise fair means without considerable effort and expenditure of time, effort, expense and/or skill. In both EU and US law, courts do not doubt that a database may be considered a trade secret. For example, the Court of Milan, Italy, specified that the commercial value of secret information does not concern the information as such, but rather, the way in which such information might be processed by 'dynamic technologies'. In fact, the Court stressed that the status of the information as a potential trade secret should be assessed in relation to the configuration and combination of the dataset as a whole and not just its individual elements in isolation. This is relevant because, after a search query, only individual elements are disclosed to the public, not the whole database.

Similar arguments are at work with the Eleventh Circuit's ruling in *Compulife.*[53] Although parts of a database can be known by the public, the database as a whole can be secret so that compilations of information are protectable even when each piece of information is under public knowledge. In the phrasing of the Restatement (Third) of Unfair Competition § 39, 'it is the

[53] *Compulife Software, Inc v Rutstein*, No. 9:16-CV-80808-JMH, 2018 WL 11033483, at *18-22 (S.D. Fla. 12 March 2018).

secrecy of the claimed trade secret as a whole that is determinative. The fact that some or all of the components of the trade secret are well-known does not preclude protection for a secret combination, compilation or integration of the individual elements'. Accordingly, a website's decision to place its database on the internet and allow public access to individual pieces of data does not mean that the website failed to take 'reasonable measures to keep [its database] secret.' The Eleventh Circuit's assumption is that the data holder had a reasonable expectation that others would not engage in massive, automated plagiarism. In other words, a trade-secret owner's failure to place a usage restriction on its website does not exclude the violation of its trade secrets.

This conclusion seems reasonable due to the state-of-the-art of technological protection measures. The ability to scrape publicly available content, register fake user accounts for malicious bots, or pass valid HTTP requests from randomly generated device IDs and IP addresses makes traditional rule-based security measures ineffective against sophisticated scraping activities. Web application firewalls are not designed to detect real-time automated threats and struggle to recognize most of today's sophisticated bots. This is why new solutions for bot detection or anti-crawler protection have been developed to identify a visitor's behavior that shows evidence of web scraping in real-time, automatically blocking malicious bots while maintaining a smooth experience for real human users. DataDome,[54] for example, compares in real-time every site hit with an in-memory pattern database to decide, through a machine learning algorithm, whether to grant access to the webpages. However, to correctly identify fraudulent traffic and block web scraping tools, a bot protection solution must be able to analyze both technical and behavioral data to set up a list of trusted partner bots, a task that may be very invasive.

Against this sort of technological cat-and-mouse game, we may thus wonder about the 'reasonable steps' that trade secret owners shall take to protect their rights. After all, in EU trade secrets litigation, one of the most common defenses is that no trade secret exists because the claimant failed to take the reasonable steps necessary to keep the information undisclosed. This defense has been successful in 59% of the instances considered in a recent empirical work.[55] Although the notion of 'reasonable steps' is generally subject to a case-specific analysis, it is arguable that – according to the EU courts – some measures that control access to and use of secret information should exist to protect such trade secrets.

Yet, in addition to the 'existential conditions' for any trade secret, we have to further consider the unlawful or improper means of web scraping to access trade secrets, namely, their misappropriation under Art. 30 of the TRIPS Agreement. The main legal issue revolves around the acquisition of such secrets by legal conduct.[56] Indeed, if data scraping activities are not unlawful as such, they may entail a lawful appropriation of the trade secret. In US law,

[54] DataDome. https://datadome.co/ (27 September 2023, date last accessed).
[55] EUIPO, Study on trade secrets litigation trends in the EU. IPR Enforcement case-law collection, 2023.
[56] WE Hilton, 'What Sort of Improper Conduct Constitutes Misappropriation of a Trade Secret' (1990) 30 Idea J L & Tech, 294-296; RG Bone, 'The Still (Shaky) Foundations of Trade Secret Law' (2014) 92 Tex L Rev,1803-1805.

the landmark case is the Fifth Circuit's decision in *E. L du PontdeNemours & Co. v. Christopher*.[57] According to the ruling, '[a]ctions may be 'improper' for trade-secret purposes even if not independently unlawful.' This means that scraping would be improper, although the activity is not per se unlawful or in breach of contract. This approach has been criticized as 'contrary to a basic understanding of trade secret law',[58] because every method of copying trade secrets would be illegal, making the requirement of the 'unlawful acquisition' either redundant or superfluous. In any event, it seems fair to affirm that, so far, applying trade secrecy law to data scraping has proved easier in the US than in EU law.

3. Discussion.

The troubles of the law with the regulation of web scraping activities offer a fruitful case of applied ethics on how to strike a balance between the pros and cons of the technology. We noted that in several cases, e.g., Ryanair's database and software, Courts have reached different, or even opposite, conclusions. Such disagreement may depend on the lack of clear rules or multiple legal cultures and jurisdictions, e.g., civil law and common law. Admittedly, there is no magic bullet to address this uncertainty, and investigations into current practices of scraping are still in progress to get out with more certainties for market operators and data subjects[59]. Yet, it seems reasonable to look for a common (and coherent) normative framework.[60] Drawing on the 'ethics of copying' mentioned in the first part of this paper, we reckon that this normative framework can be fruitfully provided by a coherent moral theory, such as the 'ethics of information'.[61] In our view, this perspective sheds light on both the troubles of the law with web scraping activities and how to ameliorate the current legal framework through the development of standards, namely, norms, values, or principles that can be adopted as the basis of a legal decision, providing thresholds of evaluation that should allow courts and policy makers to assess benefits and risks of technology.[62]

The stance of information ethics can be summed up in accordance with four laws listed to increase moral value. By assuming not only web data but also private actors, individuals, or scraping developers as 'informational entities' on the internet, the laws of information ethics recommend, on the one hand, that any kind of informational entropy – that is, any destruction or corruption of informational objects in the 'infosphere' – is evil because it entails an 'impoverishment of being' and therefore, entropy should not be caused (Law 0); or it should be prevented (Law 1); or removed (Law 2). On the other hand, Law 3 states that 'the flourishing of informational entities as well as the whole

[57] *E I du Pont deNemours & Co v Christopher*, 431 F.2d 1012, cert. denied, 400 U.S. 1024 (5th Cir. 1970).
[58] PJ Toren, 'A Dubious Decision: Eleventh Circuit Finds Scraping of Data from a Public Website Can Constitute Theft of Trade Secrets (Part I)'. IPWATCHDOG, 2 July 2020.
[59] Italian Data Protection Authority, Act 21 December 2023 [doc n 9972593].
[60] U Pagallo, M Durante, 'Three Roads to P2P Systems and Their Impact on Business Practices and Ethics' (2009) 90 J Bus Ethics, 551-564.
[61] L Floridi, *Information Ethics* (Oxford University Press 2013); M Durante, *Ethics, Law and the Politics of Information* (Springer 2017).
[62] L Busch, *Standards. Recipes for Reality* (MIT Press 2011).

infosphere ought to be promoted by preserving, cultivating, enhancing and enriching their properties'.[63] Therefore, regardless of the legal field under scrutiny with the challenges of web scraping, such challenges can be properly summed up with the laws of information ethics according to four different sets of normative issues.

The first set draws attention to all the cases in which web scraping practices arguably create or augment the 'entropy of the infosphere' by spamming email accounts, causing website crashes or setting up scams. Such malicious activities represent an 'impoverishment of the being' and typically fall under criminal law provisions. Those are the 'bad uses' of technology which should not be caused, or vice versa, should be prevented or removed.

The second set of legal issues regards controversial uses of technology that still provoke no harm. Such cases were illustrated with (i) the distinction between synchronic and diachronic extraction of data through web scraping techniques; (ii) sui generis IP rights on databases in EU law with their exceptions; down to, (iii) 'quasi-IP' regimes in US law. Insofar as data owners or data holders are not able to demonstrate that web scraping practices provoke unjust damages, the laws of information ethics on entropy do not apply since there is no harm caused or that should be prevented or removed. In some jurisdictions, e.g., EU data protection law, such harm may be presumed. We noted in section 3.6.1 that reuses of personal data for marketing purposes are lawful only on the basis of a new consent of the data subject before such reuse. Lack of consent suggests, in fact, the creation of informational entropy in the system, e.g., social and technological practices that impinge on the autonomy and self-determination of individuals. This scenario suggests why scholars often stress the limits of US data privacy law also, but not only, regarding the impact of scraping publicly available personal identifiable information.

The third set of legal issues concerns Law 3 of information ethics. There are many beneficial uses of scraping technology, such as scholars employing it for their academic research or journalists gathering and examining huge volumes of statistical data. Such uses likely contribute to the 'flourishing of informational entities as well as the whole infosphere.' Therefore, lawmakers should not only be neutral and accept those scraping practices, but rather, they should preserve, cultivate, enhance, and enrich them.

Last but not least, the fourth set of legal issues is the most controversial since it regards the hard cases of the field. The grey areas of exceptions to the general rules of the law, e.g., EU database law; the contours of the diachronic extraction of data through web scraping technology and whether such practices are at times parasitical, make it difficult to ascertain in general terms the balance that shall be struck between the pros and cons of web scraping, that is, between the 'flourishing of informational entities as well as the whole infosphere' on the one hand and, on the other, the creation of 'informational entropy.' We claim that, in such cases, lawmakers should be neutral, that is, they shall neither prevent the formation of entropy nor preserve and enhance such scraping activities. It will be up to the courts to tackle such hard cases on individual merits. Whether the doctrine of the courts turns out to be

[63] Floridi (n 61).

problematic – from the EU Court of Justice's ruling in *Innoweb* to the US Fifth Circuit's decision in *du PontdeNemours & Co. v. Christopher* – lawmakers could always intervene with their own provisions.

Conclusions.

The paper has hopefully provided an 'anatomy' of the legal challenges of web data scraping, drawing on the four laws of information ethics. The analysis has illustrated the impact of technology on current regulations, stressing the multiple, different, and even opposite opinions of courts and scholars, between civil law and common law jurisdictions, and between claims of scraping developers and web companies. The aim has been to take sides between 'good' and 'bad' uses of technology, revisiting these cases through notions of informational entropy and the flourishing of the infosphere. We stressed the four different ways in which the law should act accordingly. This normative perspective does not mean that we ended up with the magic bullet. On the contrary, attention should be drawn to all the loopholes and grey areas of the law – from the 'diachronic extraction' of web data to the 'exceptions' of EU law – which will require the development of new standards. The conjecture rests not only on disagreements among courts on the protection of databases or the parasitic nature of web scraping activities but also on further developments of technology. Scraping, as much as copying, is here to stay. The law should wisely govern its impact on the 'infosphere'.