Web Scraping in Culture and Law

Darius Rostam[*]

This article examines the practice of web scraping through its under-illuminated cultural dimension. It situates scraping within internet culture, identifying semiotic markers of both openness and enclosure. Scraping occupies an ambivalent position between these cultures, laden with symbols invoking both trespass and free access. The article analyzes how these cultural formations have historically shaped legal outcomes by identifying recurring argumentative narratives around privacy and authorization that reflect the cultural logics embedded in the law of web scraping. It demonstrates that cultural assumptions that supported a permissive legal treatment in the past do not hold for contemporary practices of AI training data collection. Yet the law has not itself enforced any particular culture – it merely demands that data hosts who wish to prevent scraping coherently signal their intent to enclose, while allowing both cultures to coexist. This antihegemonic stance proves increasingly untenable, complicated by the fact that both data hosts as well as AI companies operate as hybrids of openness and enclosure. The article advocates instead for explicit legal enforcement of either openness or enclosure.

I. Introduction

The collection of web data with automated means through web scraping has long been subject to legal controversy. After a series of scraping-friendly court decisions for web 2.0 players, web scraping law has been in constant flux and is now back in the center of attention due to the use of scraping for the compilation of artificial intelligence (AI) training data. Where much scholarship has inductively tried to bring order to the complex landscape of different legal frameworks and corresponding decisions, it is promising to look at the landscape through another lens: the under-illuminated, implicit cultural dimension of web scraping.

This article aims to make the cultural aspects of web scraping explicit and explores their influence on the law in three steps: First, it defines and explains web scraping with regards to its history and application. In a second step, the article focuses on the cultural dimension of web scraping as part of a broader internet culture. It identifies semiotic consistencies of openness and enclosure concerning the dissemination of data on the web and embeds scraping in these cultures. Lastly, the article investigates the influence of culture on the legal treatment of web scraping. This warrants a look at the relationship between culture and law in general before turning to the recurring argumentative narratives around scraping that display cultural influences,

[*] Executive Director, Center for Transnational IP, Media and Technology Law and Policy, Bucerius Law School; Dr. iur.; LL.M. (New York University). ORCID: 0000-0003-0316-1726.

and applying the findings to two current scraping scenarios, the conservation of public governmental information and the compilation of artificial intelligence (AI) training data.

The article articulates two essential points: For one, the cultural assumptions that fostered a permissive framework in the past do not hold true for current scraping efforts concerning training data for AI models. Furthermore, it is proving increasingly problematic that the law has tied the lawfulness of scraping to the notion of authorization. Rather than allowing the data host to decide whether to allow scraping, an enforcement of either openness or enclosure is necessary.

II. Web Scraping

Web scraping is an umbrella term for the automated collection of data designed to be rendered on a web page or app via the internet.[1] It involves software, known as "crawlers", "robots", or "spiders", that explores the web, retrieves relevant data from websites, and downloads it for further use.[2] Unlike using an application programming interface (API) which is designed for users to access structured data, scraping entails fetching raw data from a large number of websites, and extracting the desired elements, such as text, video, images or links.[3]

Web scraping developed as a by-product of what made the internet successfull.[4] Due to the content-agnostic technical architecture of the internet, information providers are not subject to centralized control but operate autonomously. This open yet historically contingent[5] design allows the web to accumulate a vast amount of data that is disseminated under a pull-logic: data is delivered upon client request which allows for simple protocols that lower the barrier of entry for others.[6] However, this design also makes it hard to locate information which is why as early as 1993 the first crawlers surfaced as content aggregators.[7]

There are numerous applications for web scraping. It has become increasingly important in journalism and for research data acquisition, for example concerning social media user behavior

---

[1] Megan A. Brown et al., *Web Scraping for Research: Legal, Ethical, Institutional, and Scientific Considerations*, 3 (2024), https://arxiv.org/abs/2410.23432.

[2] *See* Andrew Sellars, *Twenty Years of Web Scraping and the Computer Fraud and Abuse Act*, 24 B.U. J. SCI. & TECH. L. 372, 386 ff. (2018). *See also* Shea on Behalf of Am. Rep. v. Reno, 930 F. Supp. 916 (S.D.N.Y. 1996): "[…] software capable of automatically contacting various Web sites and extracting relevant information".

[3] *See* Jeffrey Kenneth Hirschey, *Symbiotic Relationships: Pragmatic Acceptance of Data Scraping Cyberlaw*, 29 BERKELEY TECH. L.J. 897, 897 f. (2014); Sellars, *supra* note 2 at 373.

[4] Carlos Castillo, *Effective Web Crawling*, 39 SIGIR FORUM 55, 55 (2005).

[5] DAN WIELSCH, ZUGANGSREGELN: DIE RECHTSVERFASSUNG DER WISSENSTEILUNG 249 (2008).

[6] Christopher Olston & Marc Najork, *Web Crawling*, 4 FOUNDATIONS AND TRENDS IN INFORMATION RETRIEVAL 175, 176 f. (2010).

[7] *See Id.* at 180; Seyed M. Mirtaheri et al., *A Brief History of Web Crawlers*, (2014), https://arxiv.org/abs/1405.0749.

after sites have started to restrict or discontinue access to data via APIs.[8] Most prominently, it plays a critical role for search engines to build and update their index of the web. This application was subject of a first litigation wave in which website owners challenged the scraping of online content, for example by search engines.[9] Recently, however, web scraping has again attracted attention due to its use for compiling AI model training data from the web at large scale.

III. The Cultural Dimension of Web Scraping

Web scraping has a significant yet under-illuminated cultural dimension. It is laden with semiotic code and symbols that influence its perception. Fundamentally, web scraping is embedded in internet culture. While internet culture cannot be understood as a tightly integrated, uncontested and wholly coherent world of beliefs,[10] it displays semiotic consistencies concerning the creation and dissemination of information: cultures of openness, as well as of enclosure. What makes web scraping an interesting object of cultural analysis is that is straddles both of these cultures.

1. Openness and Enclosure in Internet Culture

The internet as a communications medium is shaped by its technological architecture. A key characteristic of its design is openness:[11] The protocols on the network layer of the internet do not discriminate on a content basis,[12] which facilitates that information is generally available without restrictions on access.[13] Such openness is determined by a techno-meritocratic culture,[14] originally cultivated by the scientists that developed the net, used it for their own work and practiced free distribution and modification of source code.[15] Merit and authority in this culture of openness result not from influence of money or property rights but from reciprocal

---

[8] *See* Brown et al., *supra* note 1, at 2 (concerning X, former Twitter, and Meta). However, under Art. 40(12) of the EU Digital Services Act, researchers are given access to publicly accessible data of very large online platforms and search engines under certain conditions.

[9] *See* below IV.2.b.

[10] *See generally* William H. Sewell, *The Concept(s) of Culture*, *in* BEYOND THE CULTURAL TURN: NEW DIRECTIONS IN THE STUDY OF SOCIETY AND CULTURE 35, 57 f. (Victoria E. Bonnell & Lynn Hunt eds., 1999).

[11] MANUEL CASTELLS, THE INTERNET GALAXY: REFLECTIONS ON THE INTERNET, BUSINESS, AND SOCIETY 38 (2002); *see also* Alexander Peukert, Copyright and the Two Cultures of Online Communication (July 9, 2015), https://papers.ssrn.com/abstract=2628565 (speaking of "access"); WIELSCH, *supra* note 5.

[12] WIELSCH, *supra* note 5, at 241 ff.

[13] *See* Orin S. Kerr, *Norms of Computer Trespass*, 116 COLUM. L. REV. 1143, 1147 (2016); Patricia L. Bellia, *Defending Cyberproperty*, 79 N.Y.U. L. REV. 2164, 2192 (2004); Amber Zamora, *Making Room for Big Data: Web Scraping and an Affirmative Right to Access Publicly Available Information Online*, 12 J. BUS. ENTREPRENEURSHIP & L. 203, 225 (2019).

[14] CASTELLS, *supra* note 11, at 36 f.; another important influence on the internet's structure were military concerns and goals, *see* JANET ABBATE, INVENTING THE INTERNET 5, 145 (1999); Peukert, *supra* note 11, at 7.

[15] ABBATE, *supra* note 14, at 70 f., 73 f.; Peukert, *supra* note 11, at 7 f.

contributions to the commons.[16] Signifiers of openness, beyond the very act of serving a resource on mere request, can be found in platforms that openly distribute user generated content,[17] in meme formats that serve as semiotic templates for others to remix, in the rendering of hyperlinks as blue/purple underlined text, or simply in the ubiquitously used symbol for the internet: the globe (🌐).

Coexisting with this culture of openness, however, is a culture of enclosure.[18] As a consequence of the content-agnostic technological architecture of the internet, the decision of whether information is disseminated openly or not is delegated to its host.[19] Enclosure means that access to a resource is technically constrained and made subject to restrictive conditions such as payment, which commoditizes the underlying information. Signifiers of enclosure often mirror the logic of analogue concepts, such as metaphorical barriers like "paywalls", "watermarks", or visual obstructions, regularly introduced with a lock (🔒) or key (🔑) icon. The same applies when access to web resources discriminates by location, signified by messages such as "This content is not available in your region".

That both cultures of openness and enclosure exist side by side reflects the "thin coherence"[20] of internet culture in general. Its semiotic community shares a symbolic grammar, yet evaluates the meaning differently in moral or emotional terms.[21]

2. Web Scraping between Openness and Enclosure

Web scraping is embedded in both cultures of openness and of enclosure. Looking at how the semantic code of internet culture is made use of, it is apparent that scraping is attached with symbols of access as well as exclusivity that each ascribe, at first glance, conflicting meanings to it.

This begins with its name: To scrape, meaning "to remove something by drawing across the surface from the edge", to "scratch with finger-nails or claws", or "to collect with difficulty" has a physical, violative connotation.[22] Furthermore, metaphors such as "crawler", "spider", "robot",

---

[16] CASTELLS, *supra* note 11, at 38 ff., 48, 61.

[17] *See* Peukert, *supra* note 11, at 5.

[18] *See Id.* at 2 ff. (speaking of "exclusivity culture"); YOCHAI BENKLER, THE WEALTH OF NETWORKS: HOW SOCIAL PRODUCTION TRANSFORMS MARKETS AND FREEDOM 395 (2006) (speaking of "openness" and "enclosure"); LAWRENCE LESSIG, REMIX: MAKING ART AND COMMERCE THRIVE IN THE HYBRID ECONOMY 28 ff., 34 (2008) (distinguishing "Read/Write" and "Read/Only" culture); JULIE E. COHEN, CONFIGURING THE NETWORKED SELF 6 (2012) (speaking of "information-as-freedom" and "information-as-control").

[19] *See* WIELSCH, *supra* note 5, at 242 f. (stressing the neutrality).

[20] Sewell, *supra* note 10, at 49 f.

[21] *See generally Id.*

[22] *See Scrape, Verb*, OXFORD ENGLISH DICTIONARY, https://www.oed.com/dictionary/scrape_v (last visited Jan. 10, 2025).

or "harvester" evoke a creepy, sneaky and extractive image.[23] These terms dress a technical process of retrieving data as a violation of spaces we instinctively seek to protect.

As scraping concerns accessing resources that – although publicly accessible – are provided by others, the notion of ownership also plays an important role. Again, scraping is laden with opposing narratives: On one side, traditional symbols of space, property and enclosure are invoked when it is characterized as "stealing"[24], "theft"[25] or "trespass"[26]. Such concepts, originally applied to rival goods, imply that the requested data are tangibles subject to expropriation. Even further, imagery of "harvesting"[27] evokes ideas of harm and exploitation. In a similar vein, scraping is often contrasted with privacy,[28] a symbol of control over information against an intrusion. These associations suggest a need for protection from scraping.

On the other side stands a meaning of scraping as an act of (re)claiming public information. This understanding relies on equally traditional yet oppositional symbols of collective ownership such as the commons. Following the aphorism "information wants to be free",[29] scraping is declared as a means of innovation and societal progress or a necessary safeguard to democracy,[30] championing a public right to information. These meanings suggest that web scraping is desirable.

Lastly, another cultural axis runs along the human-machine divide. Referring to scraping software agents as "robots" or "bots" invokes the semiotics of automation. From one angle this stands for progress and modernity yet also may signify impersonality and dehumanization. It is no coincidence in this regard that bots are described with values such as "politeness"[31] or

---

[23] *See also* Timothy Erik Ström, *Data Mining on the Crawl Frontier: Metaphor in Cybernetic Capitalism*, 26 LAW TEXT CULTURE, 129 (2022).

[24] *Cf.* Daniel J. Solove & Woodrow Hartzog, The Great Scrape: The Clash Between Scraping and Privacy (July 3, 2024), https://papers.ssrn.com/abstract=4884485.

[25] Elizabeth Lopatto, *Perplexity's Grand Theft AI*, THE VERGE (June 27, 2024), https://www.theverge.com/2024/6/27/24187405/perplexity-ai-twitter-lie-plagiarism.

[26] *See* eBay, Inc. v. Bidder's Edge, Inc., 100 F.Supp.2d 1058, 1065 (2000).

[27] Solove and Hartzog, *supra* note 24, at 7.

[28] *Id.* at passim.

[29] *See Stewart Brand States Information Wants to Be Free*, GETTY IMAGES (June 15, 2012), https://www.gettyimages.in/detail/video/at-the-first-hackers-conference-in-1984-steve-wozniak-and-news-footage/146496695; R. Polk Wagner, *Information Wants to Be Free: Intellectual Property and the Mythologies of Control*, 103 COLUM. L. REV. 995 (2003).

[30] Leslie Katz, *Meet The Citizens Racing To Save Government Websites From Vanishing*, FORBES, https://www.forbes.com/sites/lesliekatz/2024/10/23/inside-the-race-to-capture-government-websites-before-they-vanish-forever/ (last visited Feb. 5, 2025).

[31] Olston and Najork, *supra* note 6, at 178; Mirtaheri et al., *supra* note 7; GitHub, *Dmi3kno/Polite*, (Nov. 17, 2024), https://github.com/dmi3kno/polite; *see also* Julie Bort, *How OpenAI's Bot Crushed This Seven-Person Company's Web Site 'like a DDoS Attack,'* TECHCRUNCH (Jan. 10, 2025), https://techcrunch.com/2025/01/10/how-openais-bot-crushed-this-seven-person-companys-web-site-like-a-ddos-attack/ (speaking of "OpenAI bot's greediness").

associated with anthropocentric concepts of "good citizens[hip]"[32], or more generally with "ethics"[33]. A bot revealing its "identity", for example, is considered good practice.[34] The dividing line between human and machine is drawn with a "CAPTCHA"[35]. It signifies the intention to exclude machines from access and therefore functions as not only a technical barrier but also as a semiotic device of a culture of enclosure.

What the semiotics of web scraping show is that it is marked by both cultures of openness and enclosure; both semiotic communities throw their symbolic net over an essentially technical process of sending requests and fetching data from web servers.[36] Scraping is therefore an arena of divergent demands for openness and enclosure, each defining different degrees of possibilities to engage in the practice. This raises the question of how law – having to reach its own determination of the meaning[37] – is positioned in this site of difference.

IV. The Influence of Culture on the Legal Dimension of Web Scraping

To investigate the influence of culture on the legal dimension of scraping, I want to first briefly frame the relationship between culture and law in general. This serves as a basis to assess in a second step whether and how law is receptive to a culture of openness and of enclosure concerning web scraping, looking both the US as well as EU legal landscape. In a third step, I scrutinize my findings by applying them to two current uses of web scraping: the conservation of public governmental information and the compilation of AI training data.

1. Culture and Law

On an abstract level, law can have very different relationships to culture. It may serve to enforce existing cultural understandings, for example when legal systems refer to *boni mores* or interpret actions under a standard of reasonableness.[38] As culture is neither stable nor coherent,[39] this often involves favoring one meaning of a sign or practice at the cost of another.[40] In contrast,

---

[32] Olston and Najork, *supra* note 6, at 178; Zamora, *supra* note 13, at 226.

[33] Brown et al., *supra* note 1, at 12, 13; Jonathan Paige, *The Legality and Ethics of Web Scraping in Archaeology*, 12 ADVANCES IN ARCHAEOLOGICAL PRACTICE 98 (2024); James Densmore, *Ethics in Web Scraping*, MEDIUM (July 23, 2019), https://towardsdatascience.com/ethics-in-web-scraping-b96b18136f01.

[34] *Cf.* Paige, *supra* note 33, at 100.

[35] A "Completely Automated Public Turing test to tell Computers and Humans Apart".

[36] *See generally* Sewell, *supra* note 10, at 51 (concerning this variant of cultural action).

[37] *See Id.* at 48 (stressing the relevance of meanings from contexts that transcend a singular social practice).

[38] *See e.g.* FRIEDRICH KARL VON SAVIGNY, OF THE VOCATION OF OUR AGE FOR LEGISLATION AND JURISPRUDENCE 17 ff. (1998); Robert Post, *Law and Cultural Conflict*, 78 CHI.-KENT L. REV. 485, 485 f. (2003) (identifying this position with *Patrick Devlin* and giving further examples from common law jurisdictions).

[39] *See* Sewell, *supra* note 10, at 52 ff.

[40] Post, *supra* note 38, at 492.

law may also have a constitutive function for culture: It can establish a culture first hand or reshape an existing culture with the consequence of possibly undermining it,[41] such as the Civil Rights Act of 1964 for instance reshaped American education culture by undermining social codes around racial segregation.

With *Post* we can imagine the intervention of law in cultural conflicts along a spectrum, depending on the controversiality of the culture in question: For seemingly stable cultural meanings, law may serve as a simple enforcer; towards contested cultural understandings within a shared culture, it gets to intervene on behalf of one side over the other; and at the end of the spectrum, for fundamental conflicts no longer about changing a shared culture, but about whether different cultures can coexist, law has to decide whether to recognize cultural pluralism or not.[42] For the resolution of such conflicts, the form of law is also relevant. There are not only rights that suppress conflict and enforce a certain understanding but also rights that promote cultural pluralism by providing an "antihegemonic domain", such as freedom of speech.[43]

2. The Culture of Web Scraping in Law

The key issue of the legal dimension of web scraping is whether information that is publicly accessible on the internet may lawfully be scraped. However, neither the US nor the EU have put in place a regulatory framework that directly addresses scraping. This underscores the relevance of the cultural dimensions of web scraping: When there are no determinative legal structuring principles readily available, the legal system is more open to be irritated by how a practice is rendered meaningful under, for example, a semiotic logic.[44]

Analyzing the US and EU legal framework reveals that the law adopts varying stances on web scraping, ranging between respecting cultural diversity and enforcing a distinctive set of cultural values. To show this, I want to forego the usual discussion of the different frameworks that regulate scraping,[45] and instead focus on the criteria the law focuses on and how they relate to cultures of openness and enclosure. While the legal treatment of scraping often is uncertain at best, recurring argumentative themes can be identified that display the cultural influences on legal decisions.

---

[41] *Id.* at 488 f.

[42] *Id.* at 493.

[43] *Id.* at 495, 501 ff.

[44] *See generally* Sewell, *supra* note 10, at 48 (concerning the autonomy of culture).

[45] *See e.g.* Hirschey, *supra* note 3, at 906 ff.; Tess Macapinlac, *The Legality of Web Scraping: A Proposal*, 71 FED. COMM. L.J. 399, 403 ff. (2018); Zamora, *supra* note 13, at 209 ff.; Han-Wei Liu, *Two Decades of Laws and Practice around Screen Scraping in the Common Law World and Its Open Banking Watershed Moment*, 30 WASH. INT'L L.J. 28, 32 ff. (2020); Maggie King, *Scraping Photographs*, 2 NOTRE DAME J. ON EMERGING TECH. 187, 193 ff. (2021).

a. Privacy

The first of such themes is privacy, which has become increasingly important due to the intensified efforts to scrape social networking sites and the enormous amount of data required for AI training.[46] Web scraping is in a tension to fundamental principles of data privacy, such as fairness, consent, transparency, data minimization, and security.[47] Public criticism arose for example in response to the scraping practices of *Clearview AI*. Clearview scraped the public web for images of faces and processed them with facial recognition algorithms to create a searchable database that featured more than 20 billion images.[48] Scraping was framed in this context as a means of "surveillance", "control"[49] and "monitoring of behavior"[50]. In the US, two lawsuits were initiated to "recapture (…) ownership"[51] and to prevent "nonconsensual harvesting"[52] of private information. These characterizations establish scraping as an intrusive and exploitative practice, suggesting a privacy-protective framework.

The law afforded means to provide such protection. Under the Biometric Information Privacy Act (BIPA) in Illinois, which was crucial in the litigation against Clearview, for the collection of biometric information a written release of the affected Illinois resident is required.[53] And under the EU General Data Protection Regulation (GDPR), obtaining individual consent (cf. Art. 6(1) lit. a GDPR) also is the only viable legal basis for web scraping because the legitimate interests of the processor (cf. Art. 6(1) lit. f GDPR) generally does not provide reliable legal justification due to the necessary balancing with fundamental rights and freedoms of the data subjects.[54]

When privacy is involved, the law therefore lends itself to enforce culture of enclosure. It does not allow for scraping of data just because it was made public for an internet audience

---

[46] *See* Solove and Hartzog, *supra* note 24, at 4, 14. *See also* the latest litigation wave of users against *Meta* concerning scraping of their personal data, e.g. BGH, Urt. v. 18.11.2024 – VI ZR 10/24, GRUR 2024, 1910.

[47] *See Id.* at 29 ff.

[48] Kashmir Hill, *The Secretive Company That Might End Privacy as We Know It*, THE NEW YORK TIMES, Jan. 18, 2020, https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

[49] OAIC, *Global Expectations of Social Media Platforms and Other Sites to Safeguard against Unlawful Data Scraping*, OAIC (Aug. 24, 2023), https://www.oaic.gov.au/news/media-centre/global-expectations-of-social-media-platforms-and-other-sites-to-safeguard-against-unlawful-data-scraping; Solove and Hartzog, *supra* note 24, at 46 ff. directly advocate for an understanding of scraping as surveillance.

[50] *See* CNIL, *Restricted Committee Deliberation No. SAN-2022-019 of 17 October 2022 concerning CLEARVIEW AI.*

[51] Sara Merken, *Clearview AI Strikes "unique" Deal to End Privacy Class Action*, REUTERS, June 13, 2024, https://www.reuters.com/legal/litigation/clearview-ai-strikes-unique-deal-end-privacy-class-action-2024-06-13/.

[52] Plaintiffs' Surreply to Defendant's Motion to Dismiss, *Am. Civil Liberties Union v. Clearview AI, Inc.*, No. 2020 CH 04353 (Ill. Cir. Ct. Jan. 5, 2021), at 2.

[53] 740 Ill. Comp. Stat. 14/15(b)(1)-(3).

[54] *See* Solove and Hartzog, *supra* note 24, at 23 f.

without explicit reservation[55] – which usually signifies a culture of openness. Instead, scraping is subjected to a default rule of enclosure. However, this must not to be understood as a reshaping of the culture of openness: The law does not ascribe a meaning to the given signifiers of openness (freely accessible data) that is contrary to their semiotic logic. Rather, the law primarily appeals to signifiers of enclosure (security,[56] ownership, and overall: privacy) *despite* recognizing the presence of said signifiers of openness.

b. Authorization

Apart from when privacy is present as a signifier, there is no enforcement of enclosure concerning publicly accessible information. The dominant theme here is authorization: The lawfulness of scraping depends on whether authorization was granted by the data host who therefore gets to decide whether scraping is allowed or not.

aa. CFAA

Authorization is directly invoked by the Computer Fraud and Abuse Act (CFAA), a federal anti-hacking statute enacted in 1984. A violation of the CFAA is given i.a. if one "intentionally accesses a computer *without authorization* or exceeds *authorized access* (…)".[57] Authorization is interpreted here constantly in analogies to trespass norms. House Reports on the CFAA, for example, analogized accessing a computer without authorization to "breaking and entering" or breaking a window and *crawling* into a home while the occupants are away.[58] Similarly, courts compare scraping to entering traditional "brick and mortar" stores.[59]

The legality of web scraping under the CFAA therefore hinges on how hosts have positioned themselves towards access requests, which is often framed in a "gates-up-or-down" inquiry: If the gate is up, scrapers have authorization, if the gate is down, access is without authorization.[60] If data is publicly available it is argued that the concept of authorization is not applicable in the first place as no gate has been erected at all.[61] For web scraping this means that under the CFAA's

---

[55] *Id.* at 41. *See also* hiQ Labs, Inc. v. LinkedIn Corporation, 31 F.4th 1180 (2022), at 11.

[56] *See Id.* at 48 ff.

[57] 18 U.S.C. § 1030(a)(2)(C), emphasis added.

[58] H.R. Rep. No. 98-894, at 20 (1984); H.R. Rep. No. 99-612, at 5 f. (1986) (emphasis added).

[59] *Cf.* eBay, Inc. v. Bidder's Edge, Inc., 100 F.Supp.2d 1058, 1065 (2000); Craigslist Inc. v. 3Taps Inc., 964 F.Supp.2d 1178, 1187 (2013); hiQ Labs, Inc. v. LinkedIn Corporation, 31 F.4th 1180, 1996 (2022); *see generally* Kerr, *supra* note 13, at 1146 (arguing in favor of the relevancy of trespass norms).

[60] *Cf. generally* Van Buren v. United States, 593 U.S. 374, 390, 141 S. Ct. 1648, 1658, 210 L. Ed. 2d 26 (2021); *see* Meta Platforms, Inc. v. Bright Data Ltd., No. 23-CV-00077-EMC, 2024 WL 251406, at *7 (N.D. Cal. Jan. 23, 2024) (concerning scraping); *Id.* at 1147, 1164 ff.

[61] hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180, 1198 f. (9th Cir. 2022); Meta Platforms, Inc. v. BrandTotal Ltd., 605 F. Supp. 3d 1218, 1262 (N.D. Cal. 2022).

criterion of authorization scraping is legal if it concerns information that is made publicly available without any conditions on access. The law therefore recognizes a culture of openness as a default.

bb. Contractual Limitations

Authorization also plays a role concerning contractual limitations on access. Many hosts try to prevent scraping under their terms of use. Courts have recently tended to enforce such contracts, allowing the hosts to prohibit scraping even of publicly available information.[62] The European Court of Justice in *Ryanair* held for instance that the freedom to use a database that is not under the protection of the Database Directive can be contractually restricted.[63] And while the 9th Circuit issued an opinion in *HiQ v. LinkedIn* that seemed to side with a culture of openness not only as a mere default but enforced even against the will of the data host,[64] *HiQ* ultimately lost on grounds of a breach of contract claim because *LinkedIn*'s user agreement expressly forbade scraping of the site. An increasingly developing exception to this are cases where the data host's terms of use are held unenforceable because they are preempted under copyright.[65] Apart from that, the lawfulness of scraping again is placed in the hands of the host that gets to define the scope of authorization via the terms of use.

cc. Copyright

Lastly, authorization is also decisive for the lawfulness of scraping under copyright law. Formative cases involve search engines that scrape web content for image search or caching purposes. Under both US law as well as the copyright law of EU member states, uses[66] of copyright protected works by search engines are generally lawful when the works were made

---

[62] *See e.g.* X Corp. v. Bright Data Ltd., 733 F. Supp. 3d 832, 837 (N.D. Cal. 2024); where breach of contract claims have failed recently, it mostly did not concern the general possibility to control scraping with contractual means, *see e.g.* Meta Platforms, Inc. v. Bright Data Ltd., No. 23-CV-00077-EMC, 2024 WL 251406, at *9 (N.D. Cal. Jan. 23, 2024) (finding that only logged-in users with active accounts were subject to the user agreements); X Corp. v. Ctr. for Countering Digital Hate, Inc., 724 F. Supp. 3d 948, 987 (N.D. Cal. 2024) (finding no foreseeability of damages).

[63] ECJ, C-30/14, ECLI:EU:C:2015:10 – Ryanair, at 39.

[64] "We agree (…) that giving companies like LinkedIn free rein to decide, (…) who can collect and use data (…) risks the possible creation of information monopolies that would disserve the public interest", hiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985, 1005 (9th Cir. 2019), cert. granted, judgment vacated, 141 S. Ct. 2752, 210 L. Ed. 2d 902 (2021).

[65] *Cf.* 17 U.S.C. § 301(a); *see e.g.* ML Genius Holdings LLC v. Google LLC, No. 20-3113, 2022 WL 710744 (2d Cir. Mar. 10, 2022); X Corp. v. Bright Data Ltd., 733 F. Supp. 3d 832, 837 (N.D. Cal. 2024).

[66] Relevant under copyright is not the act of scraping itself, but any associated reproduction of works. However, even where decisions address only the (later) distribution or making available of works, the substantial argumentation either assumes the operation-critical step of data gathering via scraping as encompassed or applies to it a maiore ad minus.

publicly available by rightholders without restrictions. This is either because the making available is treated as an authorization in form of an implied license,[67] or because the indexing is considered a fair use.[68]

The decisions emphasize the value of a default rule of openness: Search engines are regarded as necessary in the public interest for the web to be useful because they improve access to information while not being "highly exploitative".[69] The decisions are testimony of a time where the potential of the web was not fully realized yet a hope in its democratic potential justified granting search engines a broad freedom to operate, fostered by pledges of the operators to reciprocally participate in a culture of openness.[70] Furthermore, it is considered relatively easy for rightholders to instruct search engines not to scrape certain works with "widely recognized industry standards" and thereby revoke an authorization.[71] If rightholders remain silent, they are therefore considered to authorize scraping uses; if they object, however, scraping is not permitted.

c. Enforcement of Coherence, not Culture

Overall, law therefore is receptive to the cultural dimension of web scraping in two regards: If privacy is involved, the law subscribes to a culture of enclosure and enforces a corresponding set of cultural values, such as ownership, control and consent. When relying on authorization, however, law is culturally anti-hegemonic. In each of the three investigated domains, it recognizes a culture of openness in that making content publicly available without restrictions generally signifies authorization for scraping. Applied to the open architecture of the internet this represents a freedom to scrape as a default rule. Yet, this does not mean that law *enforces* a culture of openness. A consequence of authorization being the decisive criterion is that control over scraping is delegated to the data hosts. They carry the burden of taking affirmative action to revoke authorization – however, if they do so, data is enclosed and scraping becomes unlawful.

---

[67] *Cf.* Field v. Google Inc., 412 F.Supp.2d 1106, 1115 ff. (2006); Parker v. Yahoo!, Inc., No. CIV.A. 07-2757, 2008 WL 4410095 (E.D. Pa. Sept. 25, 2008); BGH, Urt. v. 29.4.2010 – I ZR 69/08, GRUR 2010, 628, 631 – Vorschaubilder I.

[68] *See e.g.* Kelly v. Arriba Soft Corp., 336 F.3d 811 (2003); *but see* Associated Press v. Meltwater U.S. Holdings, Inc., 931 F. Supp. 2d 537, 553 (S.D.N.Y. 2013) (distinguishing, however, Meltwater from a search engine).

[69] *See* Kelly v. Arriba Soft Corp., 336 F.3d 811, 818 (2003); BGH, Urt. v. 29.4.2010 – I ZR 69/08, GRUR 2010, 628, 632 – Vorschaubilder I.

[70] *See e.g. Google - About Google*, https://about.google/ (last visited Mar. 19, 2025) (stating "Our mission is to organize the world's information and make it universally accessible and useful").

[71] *Cf.* EF Cultural Travel BV v. Zefer Corp., 318 F.3d 58, 63 (1st Cir. 2003); F.3d 58 (2003); Field v. Google Inc., 412 F.Supp.2d 1106, 1116 (2006); BGH, Urt. v. 29.4.2010 – I ZR 69/08, GRUR 2010, 628, 632 – Vorschaubilder I.

Further, cultural influence becomes visible concerning the means of revoking authorization. In order to "close the gate", hosts must employ signifiers of the culture of enclosure: Revoking authorization requires either explicit means, such as natural language statements or cease-and-desist letters,[72] or implicit means such as password protection, technological barriers, or machine-readable instructions not to scrape.[73] CAPTCHAs in contrast are partly not held to be sufficient, which is consequent because they do not signify enclosure in terms of *who* is authorized to access but mark a human-machine divide on *how* content is accessed.[74]

Decisive for the law's treatment of scraping is therefore not so much the presence of signifiers of a culture of openness, but the absence of signifiers of enclosure. What is enforced is neither openness nor enclosure but their distinction: The host may prevent scraping yet has to make use of the cultural code of enclosure to do so. The law therefore allows for the coexistence of both cultures in that hosts are vested with the private power to decide how information can be used.[75] However, it prevents both cultures from clashing by demanding coherence in the host's communication of signifiers of either culture: Who wants enclosure must communicate with the corresponding signifiers.

3. Enforcement of Culture?

This antihegemonic stance of the law towards the cultures around web scraping comes at a cost. If the values of neither culture are imposed, on the one side there are hosts that generally prefer enclosure yet must carry the burden of revoking authorization, and on the other side there are scrapers that rely on the accessibility of information yet depend on the host's authorization. This provokes the question whether it is (still) adequate for the law to delegate the decision-making authority around scraping of publicly available information to its host.

Post discusses five factors to determine whether law ought to enforce certain cultural values, or should be prevented from resolving cultural conflict and instead create an environment where, like for scraping now, cultural conflict is allowed to take place:[76] First, he points to the cultural

---

[72] *See e.g.* Facebook, Inc. v. Power Ventures, Inc., 252 F. Supp. 3d 765, 768 (N.D. Cal. 2017), aff'd, 749 F. App'x 557 (9th Cir. 2019); *but see* L'Occitane, Inc. v. Zimmerman Reed LLP, No. CV 24-1103 PA (RAOX), 2024 WL 2227181, at *4 (C.D. Cal. Apr. 12, 2024).

[73] *See e.g.* EF Cultural Travel BV v. Zefer Corp., 318 F.3d 58, 63 (1st Cir. 2003) (discussing password protection); Field v. Google Inc., 412 F.Supp.2d 1106, 1113 (2006) (discussing robots.txt and meta-tags); BGH, Urt. v. 29.4.2010 – I ZR 69/08, GRUR 2010, 628, 632 – Vorschaubilder I (demanding machine-readable means); *see also* Bellia, *supra* note 13, at 2212 (suggesting a code-based approach).

[74] *Cf.* hiQ Labs, Inc. v. LinkedIn Corp., 273 F. Supp. 3d 1099, 1113 (N.D. Cal. 2017); Meta Platforms, Inc. v. Bright Data Ltd., No. 23-CV-00077-EMC, 2024 WL 251406, at *7 (N.D. Cal. Jan. 23, 2024); *see also* Kerr, *supra* note 13, at 1170.

[75] *Cf.* Peukert, *supra* note 11, at 11 ff.

[76] *See* Post, *supra* note 38, at 501 ff., 506 ff.

value to be enforced and its significance for society; second, he looks at the nature and importance of allowing for the coexistence of cultures and balances it with the advantages of imposing a certain value; third, he investigates the consequences that stem from private power being allowed to flourish in absence of dispositive legal rules; fourth, he points to the costs of enforcement of cultural values; and fifth, he takes the cost of persistent cultural disagreement into account.[77]

Applying these factors to two examples from the recent past, the conservation of public governmental information and the compilation of AI training data, shows that the reluctance of the law to enforce either openness or enclosure over the other is no longer sustainable for today's web scraping applications.

a. Conservation of Public Governmental Information

A large amount of publicly accessible data on the web is provided by governments for whom the web serves as the main channel to share information with the general public.[78] However, compared to the analogue age, digital distribution allows for larger control over publicized information. For this reason, there have been efforts by citizens to scrape and save US government websites, for example at the end of presidential administrations since at least 2008.[79]

The importance of this practice was emphasized when the first Trump administration removed and altered environmental data on websites of US federal environmental agencies,[80] prompting efforts from scientists to preserve the unmodified data.[81] Most recently in early 2025, the second Trump administration began to purge certain terms and datasets from websites of the federal government on a grand scale,[82] which again led to efforts to scrape and restore the concerned data.

---

[77] *Id.* at 506 ff.

[78] *Cf.* Julian Lucas, *The Volunteer Data Hoarders Resisting Trump's Purge*, THE NEW YORKER, Mar. 2025, https://www.newyorker.com/news/the-lede/the-data-hoarders-resisting-trumps-purge.

[79] *See e.g. End of Term Web Archive*, END OF TERM WEB ARCHIVE, https://eotarchive.org/ (last visited Mar. 19, 2025).

[80] *See* Eric Nost et al., *Visualizing Changes to US Federal Environmental Agency Websites, 2016–2020*, 16 PLOS ONE e0246450 (2021) (pointing out that the use of the term "climate change" i.a. decreased by an estimated 38%, and access to as much as 20% of the Environmental Protection Agency's website was removed).

[81] *See* Justine Calma, *How Scientists Scrambled to Stop Donald Trump's EPA from Wiping out Climate Data*, THE VERGE (Mar. 8, 2021), https://www.theverge.com/22313763/scientists-climate-change-data-rescue-donald-trump.

[82] *See e.g.* Will Stone, *Trump Administration Purges Websites across Federal Health Agencies*, NPR, Jan. 31, 2025, https://www.npr.org/sections/shots-health-news/2025/01/31/nx-s1-5282274/trump-administration-purges-health-websites; Justine Calma, *Donald Trump's Data Purge Has Begun*, THE VERGE (Jan. 31, 2025), https://www.theverge.com/news/604484/donald-trumps-data-purge-has-begun.

Applying *Post's* factors in this context shows that it is misguided to tie the lawfulness of scraping of public governmental information to an authorization by the government:[83] First, the significance of the cultural value to be enforced, the open access to information for the purpose of conservation, is substantial. A democratic, bottom-up decision-making process cannot take place if access to information is controlled top-down. Second, the enclosure of public information does not serve to realize important cultural norms as there is no legitimate governmental interest in preventing the scraping of already accessible information that the general public paid for. Third, it is the openness of information that would precisely prevent private actors from dominating the public discourse. Fourth, the effects of enforcing a culture of openness are negligible because the information will be produced either way and scraping does not diminish its value or availability. Finally, allowing the government to decide whether information can be scraped is costly as it opens the door for selective intransparency, which in the long run leads to an erosion of trust. There is historic precedent in this regard when it was discovered in 2003 that the Bush White House had selectively blocked search engines from scraping controversial content related to the Second Gulf War.[84]

b. Compilation of AI Training Data

Web scraping has also come back into the spotlight because of the compilation of AI training data. Under the hypothesis that a model's cognitive ability scales with an increase in dataset size,[85] AI companies strive to compile ever larger training datasets. While some have stopped making the composition of their training datasets transparent,[86] it is evident that they in large part consist of content scraped from the web.[87] Web content is well suited for AI training

---

[83] As government works are in general not subject to copyright protection (cf. 15 U.S.C. § 105(a)), restrictions to scraping could result only from violations of the CFAA or contractual terms.

[84] *See White House Says Blocking Iraq Web Documents Was "Mistake,"* CNET, https://www.cnet.com/tech/tech-industry/white-house-says-blocking-iraq-web-documents-was-mistake/ (last visited Mar. 20, 2025).

[85] *See e.g.* Bruce Rushing & Javier Gomez-Lavin, *Is the Scaling Hypothesis Falsifiable?*, (2024), https://philsci-archive.pitt.edu/23622/.

[86] *See e.g.* Tom B. Brown et al., *Language Models Are Few-Shot Learners*, 9 (2020), https://arxiv.org/abs/2005.14165 (mentioning for GPT-3 i.a. Common Crawl, WebText2, Books1, Books2 and Wikipedia, which rely on web scraping); Hugo Touvron et al., *LLaMA: Open and Efficient Foundation Language Models*, (2023), https://arxiv.org/abs/2302.13971 (naming for LLaMA of Meta AI Common Crawl, Github, Wikipedia, Stack Exchange, ArXiv, Project Gutenberg and Books3).

[87] *See Our Approach to Data and AI*, OPENAI, https://openai.com/index/approach-to-data-and-ai/ (last visited Jan. 30, 2026) (claiming to take web crawler permissions for AI into account each time a new model is trained); Billy Perrigo, *TIME100 AI 2024: Sundar Pichai*, TIME (Sept. 5, 2024), https://time.com/7012748/sundar-pichai-2/ (stressing that Google has "given a clear way for people to opt out of [...] AI training"); Jess Weatherbed, *Anthropic's Crawler Is Ignoring Websites' Anti-AI Scraping Policies*, THE VERGE (July 25, 2024), https://www.theverge.com/2024/7/25/24205943/anthropic-ai-web-crawler-claudebot-ifixit-scraping-training-data (observing Anthropic's Claude Bot to engage in comprehensive scraping); *Introducing Meta Llama 3*, META AI, https://ai.meta.com/blog/meta-llama-3/ (last visited Jan. 30, 2026) (stating that "Llama 3 is pretrained on over 15T tokens that were all collected from publicly available sources").

due to its diversity in viewpoints, styles, and media types, which has a positive effect on model performance. It also represents a low-hanging fruit because it is cheap to collect by making use of publicly available datasets. And while it appears that AI companies have exhausted the initial potential of web data and are increasingly seeking data through licensing deals or digitization efforts,[88] web scraping remains relevant to keep existing datasets up to date, and to satisfy the user demand for timely information.

aa. Hybrids of Openness and Enclosure

The compilation of training data through scraping is met with resistance from the "scraped" parties that host or assert ownership over the compiled data. The conflict has a significant cultural dimension because the scraped data is used for training, for example, generative AI models which stands in tension with established norms around how cultural artifacts were leveraged before in service of innovation, such as attribution, permission, and compensation. Against this the scraped parties invoke familiar symbols of a culture of enclosure, basing legal claims on ownership and consent.[89] The scrapers in contrast conceive of scraping as a means for exercising a freedom to "learn" from public information as part of a commons in the interest of innovation.[90] They invoke a culture of openness.

What further complicates the conflict, however, is that both sides are hybrids; they operate under the logic of not just one, but both cultures simultaneously.[91] The scrapers on the one side invoke openness but on the other side practice enclosure concerning the products they build based on scraped public data. The models of the leading AI firms are proprietary,[92] access is subject to permission and payment, and restrictive terms of use hinder users from learning from the models in turn.[93] This marks a major cultural difference to the first wave of lawsuits around

---

[88] *See e.g. Partnership with Axel Springer to Deepen Beneficial Use of AI in Journalism*, OPENAI (Mar. 13, 2024), https://openai.com/index/axel-springer-partnership/.

[89] *See e.g. Authors Guild Open Letter to Generative AI Leaders*, https://actionnetwork.org/petitions/authors-guild-open-letter-to-generative-ai-leaders (last visited Mar. 21, 2025); Lopatto, *supra* note 25.

[90] *See e.g. OpenAI's Proposals for the U.S. AI Action Plan*, OPENAI, https://openai.com/global-affairs/openai-proposals-for-the-us-ai-action-plan/ (last visited Mar. 20, 2025); *Google's Comments on the U.S. AI Action Plan*, GOOGLE (Mar. 13, 2025), https://blog.google/outreach-initiatives/public-policy/google-us-ai-action-plan-comments/.

[91] *See* LESSIG, *supra* note 18, at 179 (describing hybrids as a third economy built upon sharing and commercial economies); *see also* Peukert, *supra* note 11, at 11 ff.

[92] "Open" AI systems are in this regard a misnomer, *see* David Gray Widder, Meredith Whittaker & Sarah Myers West, *Why 'Open' AI Systems Are Actually Closed, and Why This Matters*, 635 NATURE 827 (2024).

[93] *Cf.* Jason Koebler, *OpenAI Furious DeepSeek Might Have Stolen All the Data OpenAI Stole From Us*, 404 MEDIA (Jan. 29, 2025), https://www.404media.co/openai-furious-deepseek-might-have-stolen-all-the-data-openai-stole-from-us/; *see also* Mark A. Lemley & Peter Henderson, The Mirage of Artificial Intelligence Terms of Use Restrictions (Dec. 9, 2024), https://papers.ssrn.com/abstract=5049562 (questioning the legal enforceability of such terms).

scraping for web search. Past decisions were marked by the often implicit assumption that such applications would provide for reciprocal openness.[94] The scraping AI companies, however, keep their models enclosed and produce a substitute instead of a complement, allowing so far for no direct participation of the data hosts in the created value so far.[95] The scraped data hosts on the other hand invoke the semantics of enclosure yet rely on signifiers of the culture of openness by making their information publicly available without any reservations. Even if they disapprove of the scraping of their data, they still want to take advantage of the discoverability of publicly available information on the web.

bb. Favoring Congruence Over Pluralism

When applying *Post's* factors to decide between a logic of cultural congruence or an anti-hegemony under the law, here the question is not so much whether law should enforce openness over enclosure or vice versa. The issue concerning scraping is rather whether the enforcement of *either one* culture is preferable to the current antihegemonic stance, where the data host is vested with the decision-making authority on the lawfulness of scraping.

First, both positive cultural values at stake may claim some importance to society. While the recent advancements in AI model capabilities would likely not have been achieved but for scraping of the open web, the enclosure of information, at least to enforce a compensation, seems equally relevant to sustain the environment that provided for this innovation in the first place.

The second factor focuses on the cultural norm realized under the status quo, here the self-determination of the data hosts. At first glance, this seems to allow for both openness and enclosure without incurring the cost of suppressing either: The default is openness, and where the provision of data requires a compensation, the host may enclose information and enforce it. However, there is hardly any way to effectively prevent the scraping of publicly available information on the web without also removing it from its attention economy. And while powerful information hosts have the means to protect their information and market it even

---

[94] Search engines, for example, use scraped thumbnails in their image search to direct traffic to the source website. This assumption underlies several decisions, *see e.g.* Associated Press v. Meltwater U.S. Holdings, Inc., 931 F. Supp. 2d 537, 553 (S.D.N.Y. 2013) (distinguishing between complementary search engines and substituting news clipping on the grounds of a low click-through rate); Kelly v. Arriba Soft Corp., 336 F.3d 811, 819 (2003); Field v. Google Inc., 412 F.Supp.2d 1106, 1114 (2006). *See also* above IV.2.b.cc.

[95] This is illustrated by increasing complaints of website providers that rely on traffic from search engines yet now with AI generated search responses experience a downturn, *see* Jody Godoy, *Google's AI Previews Erode the Internet, US Edtech Company Says in Lawsuit*, REUTERS, Feb. 25, 2025, https://www.reuters.com/legal/googles-ai-previews-erode-internet-edtech-company-says-lawsuit-2025-02-24/.

behind paywalls,[96] small hosts depend on visibility and are therefore often forced to expose themselves to the risk of being scraped. The disadvantages of self-determination therefore likely do not outweigh the advantages of enforcing either openness or enclosure.

Concerning the third factor, the case for changing the status quo to either openness or enclosure is also strong. Under the current scraping regime, the leading AI companies have a first-mover advantage. Subsequent companies are no longer able to collect the same wealth of data because, due to rising public awareness for the issue, significantly more hosts have now objected to scraping.[97] Maintaining the status quo would therefore mean to erect a meaningful barrier to entry to a nascent market, allowing private players to dominate a key technology.

The fourth factor takes the negative effects of enforcing either openness or enclosure into account. Under an enclosure regime, scraping will most likely become impractical for many applications because consent-based access would have to overcome vast transaction costs. A strict openness regime in contrast would fare no better because uses that substitute for the original provision of information can proceed without compensation. However, openness has the advantage that it can be coupled with further conditions, such as a compensation regime. This solution is superior to the status quo which for most data hosts consists of a binary between an unconditional allowing or disallowing.

Lastly, the analysis points to the disadvantages of the current coexistence of openness and enclosure at the discretion of the data host. In lieu of legal protection, it is up to the host to prevent scraping which introduces wasteful costs for the development of technical solutions better addressed by an enforceable legal rule.[98] Far worse, however, the culturally pluralist status quo has already led to a significant deprivation of public web data as more and more hosts revoke their authorization for scraping in general, not only for AI training data purposes.[99] The cultural conflict around AI training therefore spreads to other domains with completely different interests and bargains, such as web archiving. It shows that the technical means of expressing authorization are not granular enough to allow for a sustainable coexistence of openness and enclosure in other domains.

Overall, the factors clearly identify the deficiencies of the anti-hegemonic status quo when compared with an enforcement of either openness or enclosure. They also show that while both

---

[96] The big players are also able to leverage enclosure in order to negotiate deals for scraping uses, *see e.g.* Partnership with Axel Springer to deepen beneficial use of AI in journalism, *supra* note 88.

[97] *See* Shayne Longpre et al., Consent in Crisis: The Rapid Decline of the AI Data Commons (July 24, 2024), http://arxiv.org/abs/2407.14933 (observing a rapid crescendo of data restrictions from web sources).

[98] *See* Solove and Hartzog, *supra* note 24, at 22 (pointing to the developing "technological front").

[99] *See* Longpre et al., *supra* note 97.

cultures carry their cost with them, a regime of openness coupled with a compensation regime promises a superior solution for the conflict around web scraping for AI training purposes.

V. Conclusion

The article has explored the cultural and legal dimensions of web scraping. As part of a thinly coherent internet culture, scraping is embedded in both a culture of openness as well as enclosure, which define different degrees of possibilities to scrape. In reaching its own determination of the meaning, the law is influenced by these cultural values: While the presence of the signifier privacy leads to an enforcement of enclosure of publicly available information, mainly, the law recurs to the notion of authorization by the data host. It therefore delegates the decision of whether scraping may lawfully take place to the data host and does neither enforce a culture of openness nor enclosure but only the coherence of communicating in either of them when making information publicly available.

This proves problematic concerning two current fields of application of scraping: On the one hand, recent efforts to conserve public governmental information show that scraping must also be possible without authorization. Otherwise, a government could influence the democratic process. On the other hand, the compilation of AI training data demonstrates the advantages of enforcing either culture around scraping, preferably openness coupled with further means of compensation, to avoid the disadvantages of the current cultural pluralism. The analysis also shows that some cultural assumptions that marked permissive decisions in the past, such as a culture of reciprocal openness cultivated by the scraper, no longer hold true for the current scraping efforts around AI training data.

Web scraping therefore is at a crossroads: The law must evolve beyond the binary of authorization to address the nuanced needs of our information ecosystem. Otherwise, both legitimate interests in enclosure as well as innovation brought about by openness will suffer.